rocha@lanl.gov

Extraction and Semi-metric Analysis of Social and Biological Networks

Luis M. Rocha
Complex Systems Modeling
CCS3 - Modeling, Algorithms, and Informatics
Los Alamos National Laboratory, MS B256
Los Alamos, NM 87545
and
Center for Computational Biology
Instituto Gulbenkian de Ciência
Oeiras, Portugal

NASA's Goddard Space Flight Center
Information Science & Technology (IS&T) Colloquium
January 29, 2003

Los Alamos
National Laboratory

Luis Rocha
2002

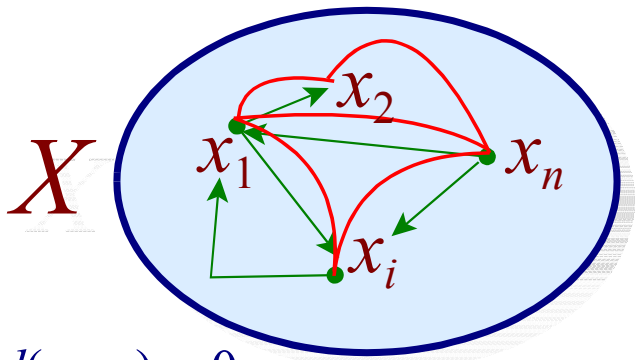
http://www.c3.lanl.gov/~rocha

rocha@lanl.gov

Distances on Graphs

Measured from associative "knowledge" graphs

d is a distance function on set *X* if it is a nonnegative, symmetric, real-valued function such that $d(x, x) = 0$ (Shore & Sawyer 1993)



$d(x_i, x_i) = 0$

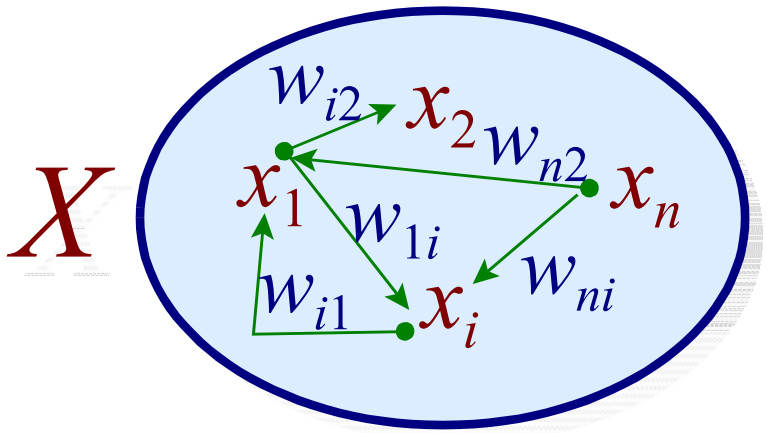
$d(x_i, x_j) = 1$, if there is an edge

$d(x_i, x_k) = d(x_i, x_j) + ... + d(x_l, x_k)$ 1, if there is a path

Due to the symmetry requirement, distance functions yield non-directed distance graphs

$$d(x_1, x_2) \leq d(x_1, x_3) + d(x_3, x_2)$$

Metric: the smallest distance between nodes is always the most direct path



In real-valued weighted graphs, derived distance functions can be semi-metric

$$d(k_1, k_2) > d(k_1, k_3) + d(k_3, k_2)$$

Semi-metric

In graphs used to store "knowledge", what does it mean?

Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

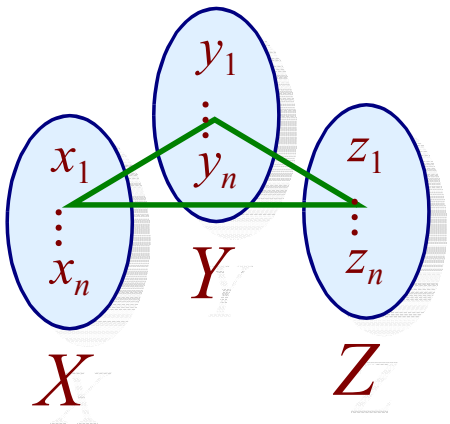
Los Alamos National Laboratory

rocha@lanl.gov

Mathematical Background

Relations

Represent the presence or absence of association, interaction or interconnectedness between the elements of two or more sets.



$r(x_i, y_j, z_k) = 1$

- a relation R between sets $X_1, X_2, ..., X_n$ is a subset of the Cartesian product of these sets: $R(X_1, X_2, ..., X_n) \subseteq X_1 \times X_2 \times ... \times X_n$.
  - Traditional logical operations between sets can be used to modify relations

$$R(x_1, x_2, ..., x_n) = \begin{cases} 1 & \text{iff}(x_1, x_2, ..., x_n) \in R \\ 0 & \text{otherwise} \end{cases}$$

$R(\mathbf{x}) \in [0,1], \quad \forall \mathbf{x} \in \mathbf{X}$ Fuzzy: Degree of Relation or association

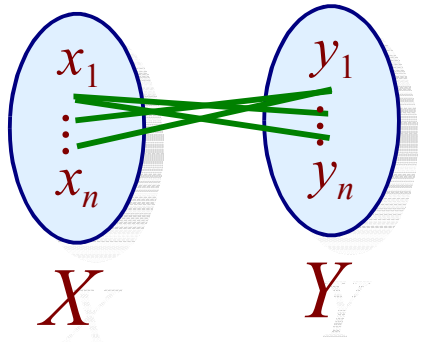
Luis Rocha
2002

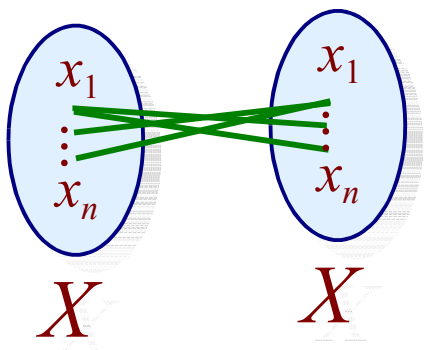
http://www.c3.lanl.gov/~rocha

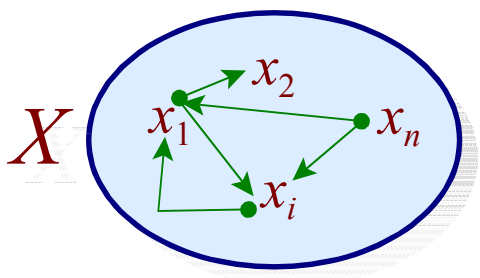
Los Alamos National Laboratory

rocha@lanl.gov

Binary Relations







- Binary fuzzy relations are a generalization of real functions
  - Two or more elements of Y may relate to an element of X
  - Easily represented by matrices of dimension $n \times m$

- Graphs are binary relations defined on a single set: $R(X, X)$.
  - Degrees of association between elements of the same set
  - If symmetric, R represents a non-directed graph

Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Fuzzy Graphs

Properties

- Reflexive
  - iff $R(x, x) = 1$ for all $x \in X$
    - every element of X is maximally associated with itself
- Symmetric
  - iff $R(x, y) = R(y, x)$ for all $x, y \in X$
    - Matrices require only $(n^2\text{-}n)/2$ elements to be defined
- (Max-Min) Transitive
  - iff $R(x, z) \geq \max_{y \in X} \min[R(x, y), R(y, z)]$ for all $x, z \in X$
    - For each indirect connection between x and z through some y, the weight of the connection is the smallest of each connection (x to y and y to z). Finally, the weight of the connection between x and z, is the largest of all indirect connections through all y (strongest path defined by weakest link)

rocha@lanl.gov

Composition of Fuzzy Graphs

Max-Min Composition: $R \circ R = \max_k \min(r_{ik}, r_{kj}) = r'_{ij}$

where r_{ij} denotes $R(x_i, x_j)$

The max-min (logical) composition of matrices is performed in the same way as the numerical counterpart, except that *multiplication* and *summation* are substituted by the *Min* (and) and *Max* (or) operations respectively.

- Transitive closure of a relation $R(X, X)$
  - The relation that is transitive, contains $R(X, X)$, and whose elements have the smallest possible membership weights that still allow the first two requirements.
    - It yields a relation where all pairs of elements which were directly or indirectly related in the original relation, are now directly related
    - 1. $R' = R \cup (R \circ R)$; 2. If $R' \neq R$, make $R = R'$ and go back to step 1; 3. Stop: $R_T = R'$

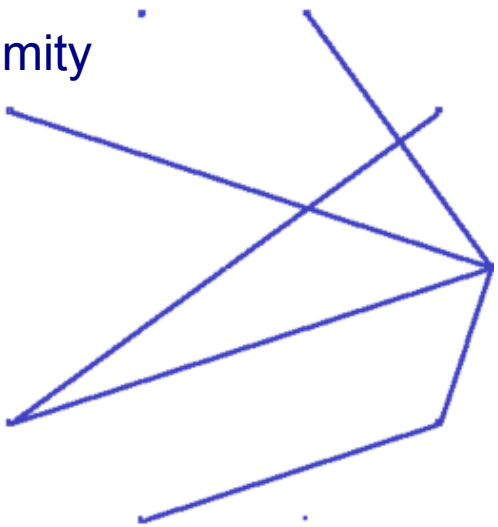
Luis Rocha
2002

rocha@lanl.gov

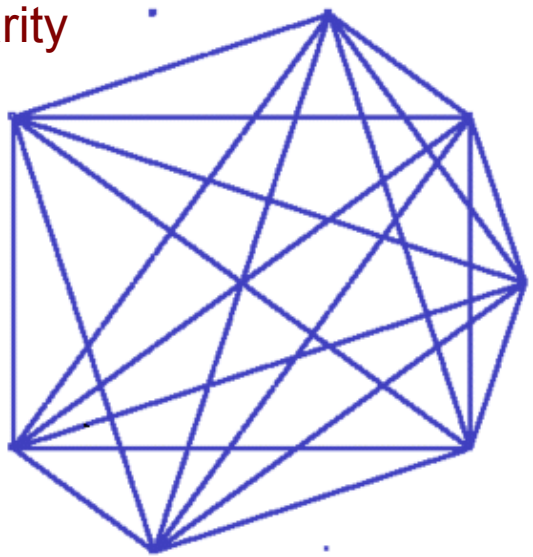
Similarity and Proximity Relations

- Similarity Relation
  - A reflexive, symmetric, and transitive binary fuzzy relation
    - Also known as an equivalence relation.
- Proximity Relation
  - A reflexive and symmetric binary fuzzy relation
    - Also known as a compatibility relation
    - The transitive closure of a proximity relation is a similarity relation.

Proximity



Similarity



Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos
National Laboratory

rocha@lanl.gov

Measuring "Knowledge" Associations

From Document Relations

- Document × Document
  - Co-Citation or Hyperlink structure
- Document × Keyterms
  - Keyterm Co-Occurrence
- Document/Dataset × Gene Expression
  - Gene Co-Occurrence or Co-Expression
- Document × Author
  - Co-Authorship (Collaboration Network)

Given a binary relation R between sets X and Y we extract two proximity relations: $XYP(x_i, x_j)$ is the probability that both x_i and x_j are related in R to the same element $y \in Y$. Conversely, $YXP(y_i, y_j)$ is the probability that both y_i *and* y_j are related in R to the same element $x \in X$.

$$XYP(x_i, x_j) = \frac{\sum_{k=1}^{m} (r_{i,k} \wedge r_{j,k})}{\sum_{k=1}^{m} (r_{i,k} \vee r_{j,k})}; \quad YXP(y_i, y_j) = \frac{\sum_{k=1}^{n} (r_{k,i} \wedge r_{k,j})}{\sum_{k=1}^{n} (r_{k,i} \vee r_{k,j})}$$

Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Document and Keyword Proximity

Typical IR Example: Knowledge Implied by Keywords

Given a binary relation A between sets of keywords K and documents D we extract two proximity relations: $KDP(k_i, k_j)$ is the probability that both keywords k_i and k_j co-occur in the same document $d \in D$. Conversely, $DKP(y_i, y_j)$ is the probability that both documents d_i *and* d_j contain the same keyword $k \in K$.

$$kdp(k_i,k_j)=\frac{\sum_{k=1}^{m}(a_{i,k}\wedge a_{j,k})}{\sum_{k=1}^{m}(a_{i,k}\vee a_{j,k})}=\frac{N_{\cap}(k_i,k_j)}{N_{\cup}(k_i,k_j)}$$

(Keyword Document Proximity)

$$dkp(k_i,k_j)=\frac{\sum_{k=1}^{m}(a_{i,k}\wedge a_{j,k})}{\sum_{k=1}^{m}(a_{i,k}\vee a_{j,k})}=\frac{N_{\cap}(d_i,d_j)}{N_{\cup}(d_i,d_j)}$$

(Document Keyword Proximity)

rocha@lanl.gov

Pointwise Mutual Information

Other non-proximity Probability Measures of Association

Conditional Probability

Given a binary relation R between sets X and Y we extract :

$$P_X\left(x_i, x_j\right) = P\left(x_i \middle| x_j\right) = \frac{\sum_{k=1}^{m}\left(r_{i,k} \wedge r_{j,k}\right)}{\sum_{k=1}^{m}\left(r_{j,k}\right)}; \quad P_Y\left(y_i, y_j\right) = P\left(y_i \middle| y_j\right) = \frac{\sum_{k=1}^{n}\left(r_{k,i} \wedge r_{k,j}\right)}{\sum_{k=1}^{n}\left(r_{k,j}\right)}$$

$P_X(x_i, x_j)$ is the probability that x_i is related in R to $y \in Y$, given that x_j is related to y.

$P_Y(y_i, y_j)$ is the probability that y_i is related in R to $x \in X$, given that x_y is related to x.

rocha@lanl.gov

Proximity vs. Conditional Probability

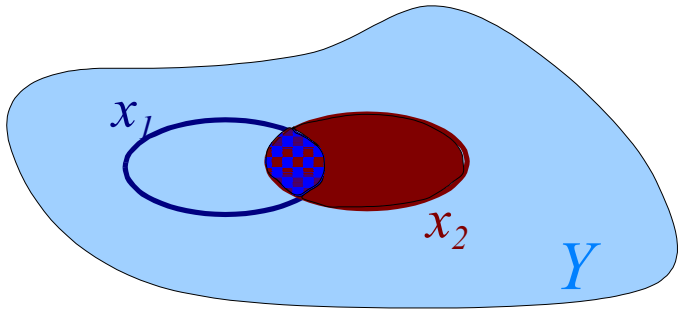
$$XYP(x_i, x_j) = \frac{1}{\frac{1}{P_X(x_j, x_i)} + \frac{1}{P_X(x_i, x_j)} - 1}; \quad YXP(y_i, y_j) = \frac{1}{\frac{1}{P_Y(y_j, y_i)} + \frac{1}{P_Y(y_i, y_j)} - 1}$$

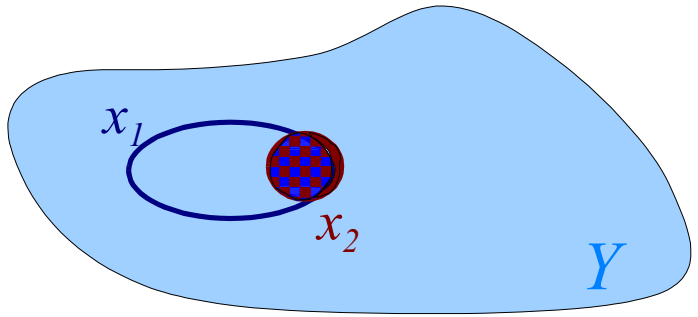
- P_X and P_Y are not symmetric
  - can measure a strong degree of association between two elements, when that association is one-sided only.
  - In many applications, when we think of a strong association between two elements, we expect both directions of association to be similar.
  - Distances are symmetric, proximity is the semantic inverse of distance.

rocha@lanl.gov

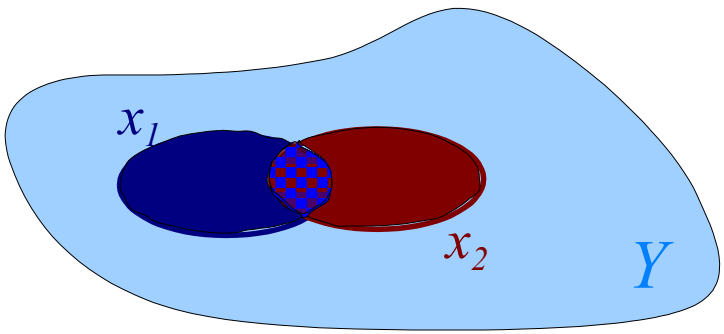
Conditional Probability and Proximity Measures

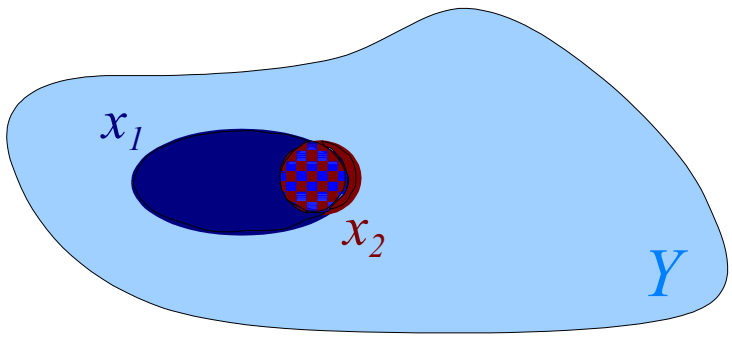
Conditional Probability





Proximity





Luis Rocha
2002

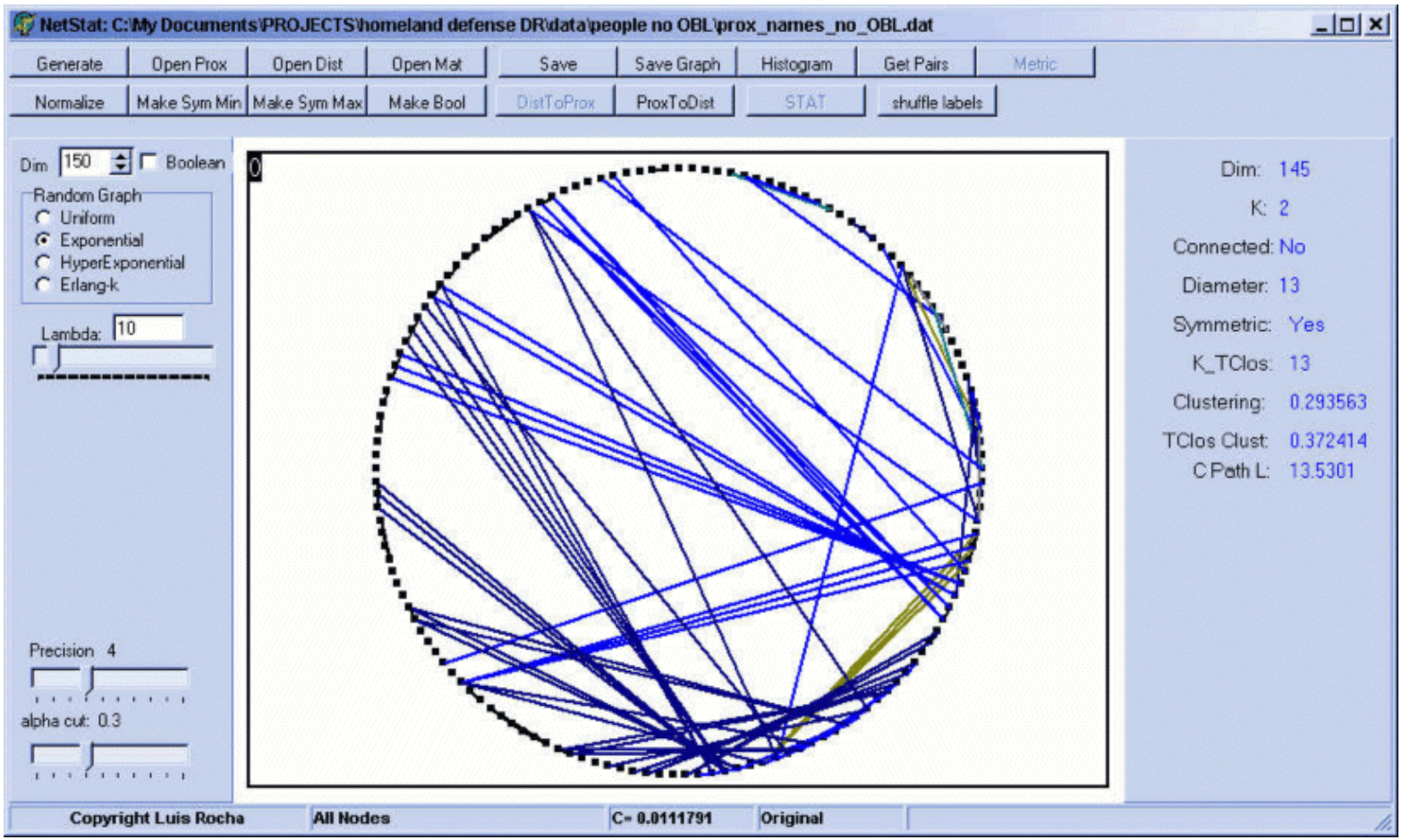
http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Social Proximity Networks

Terrorist Associations: People Document Proximity (PDP)



Luis Rocha
2002

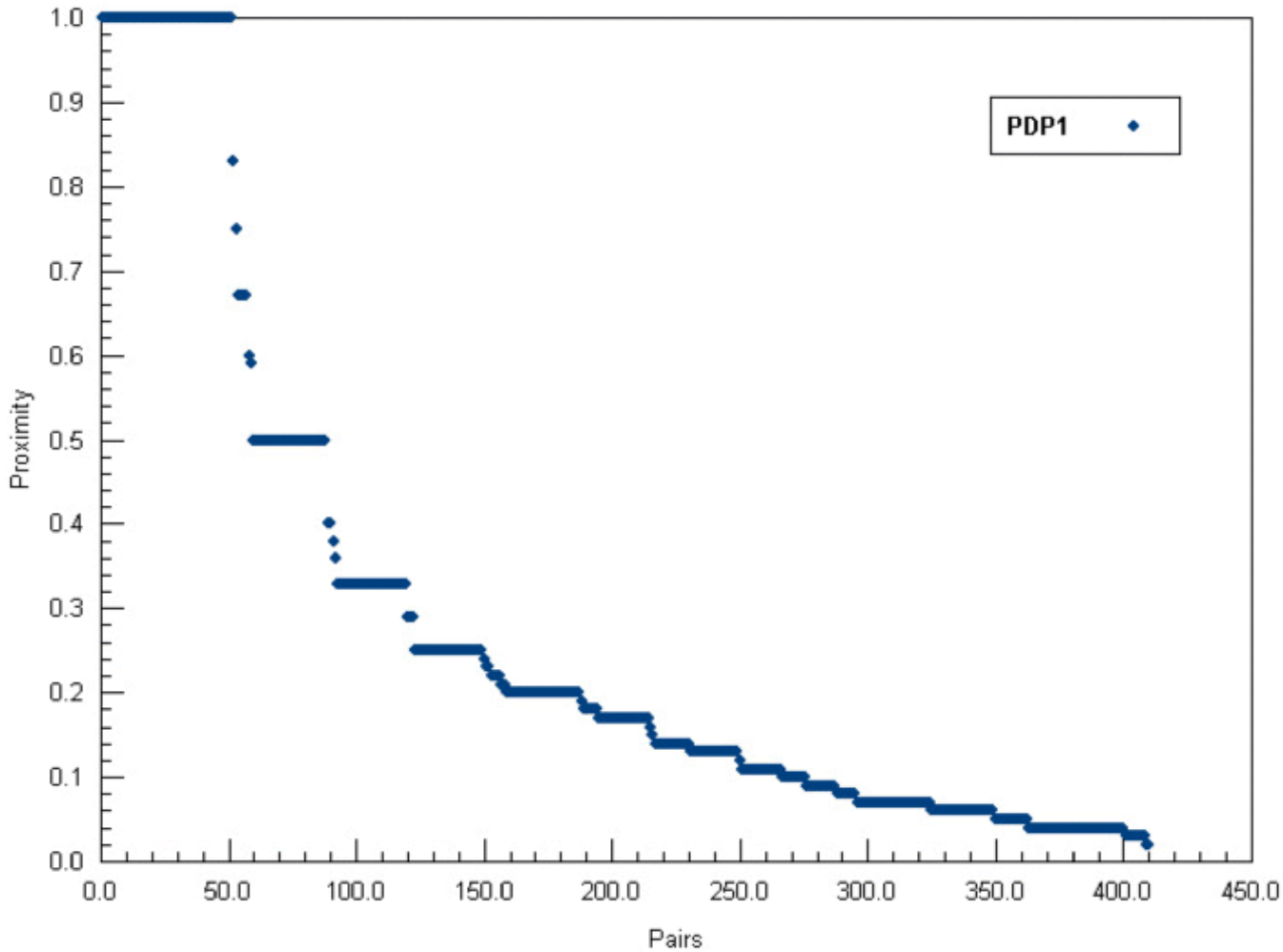
http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Proximity Distribution

PDP: exponential distribution



Luis Rocha
2002

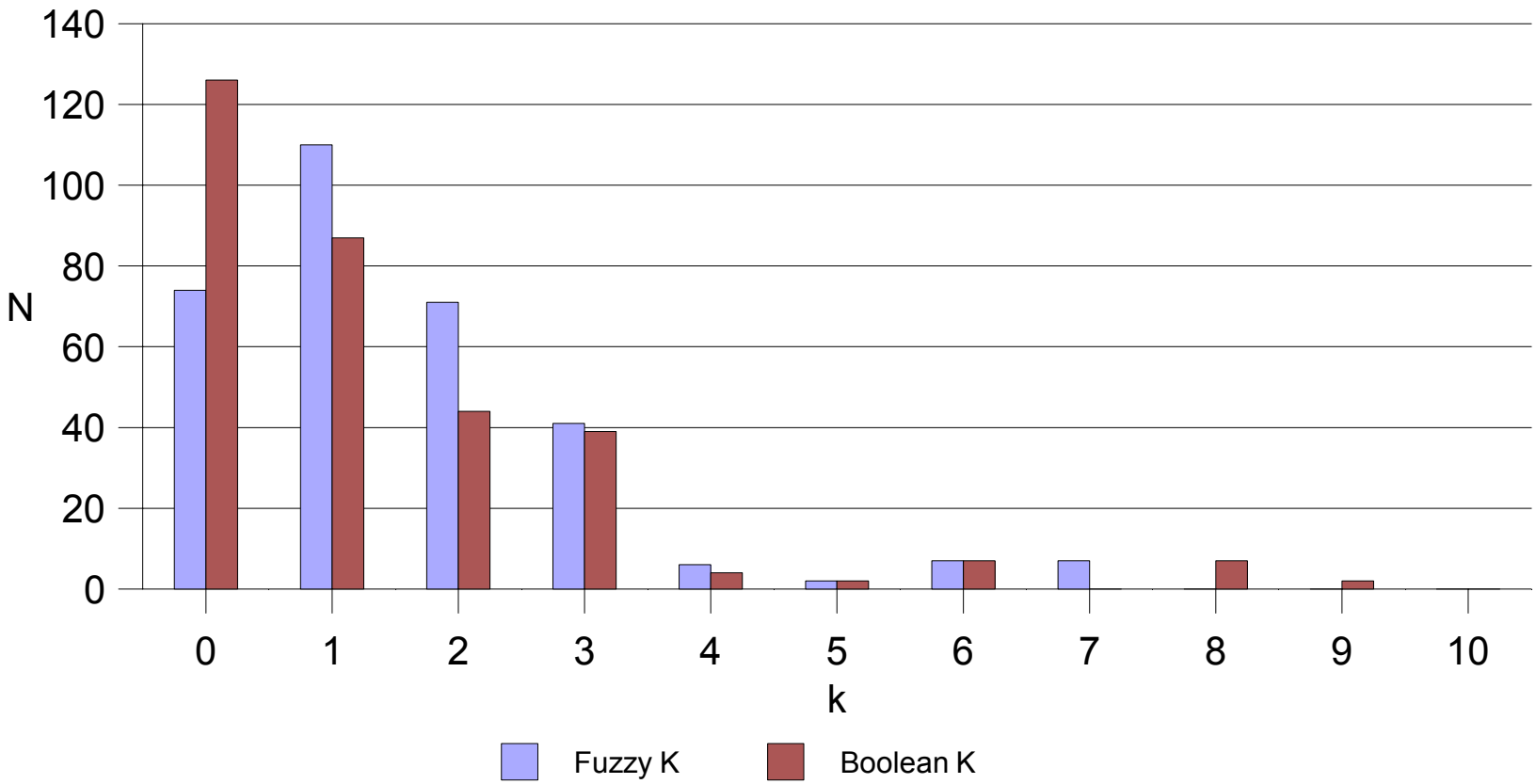
http://www.c3.lanl.gov/~rocha

Los Alamos
National Laboratory

rocha@lanl.gov

PDP Histogram fo Node Degree

318 People Names



Luis Rocha
2002

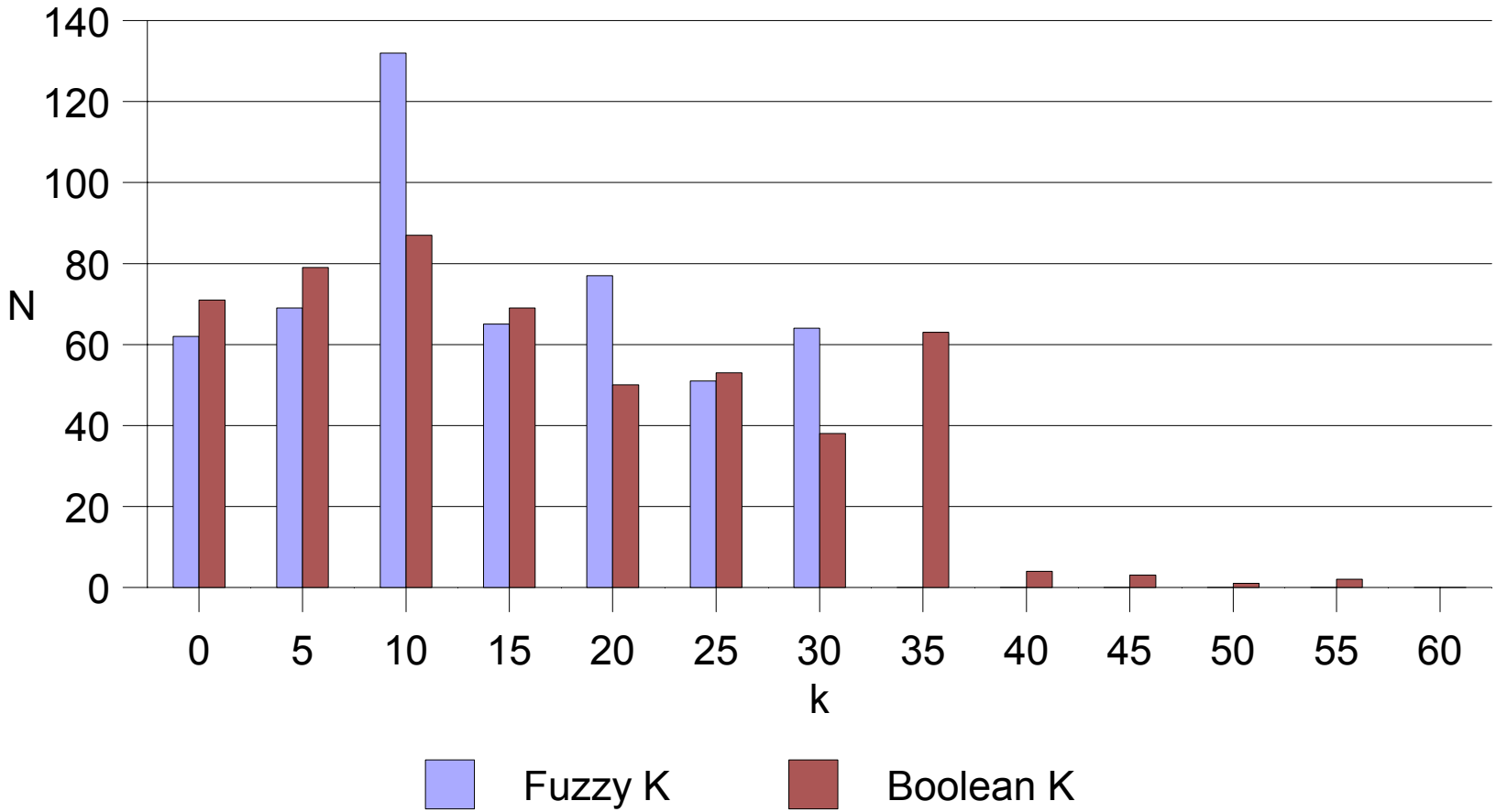
http://www.c3.lanl.gov/~rocha

Los Alamos
National Laboratory

rocha@lanl.gov

IPP Histogram of Node Degree

520 ISSN



Luis Rocha
2002

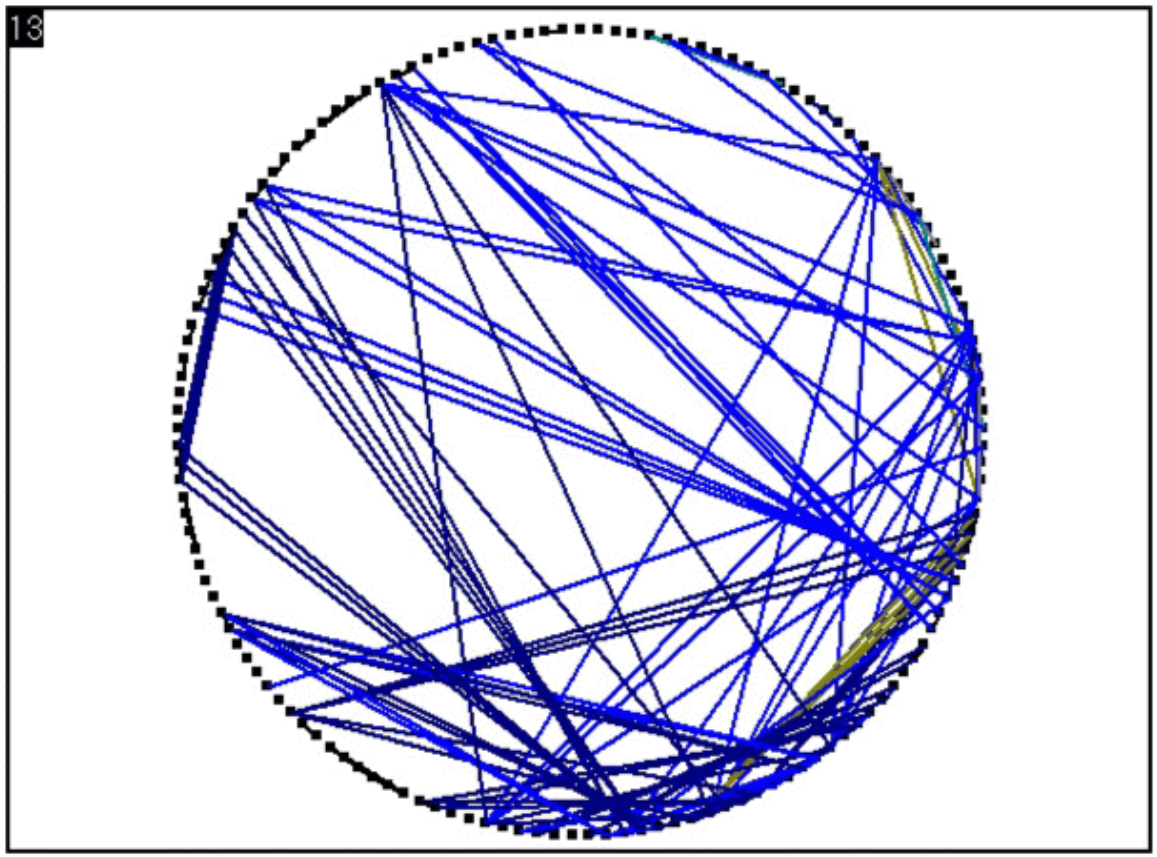
http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Terrorist Networks

PDP: transitive Closure



Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos
National Laboratory

rocha@lanl.gov

Terrorist Networks

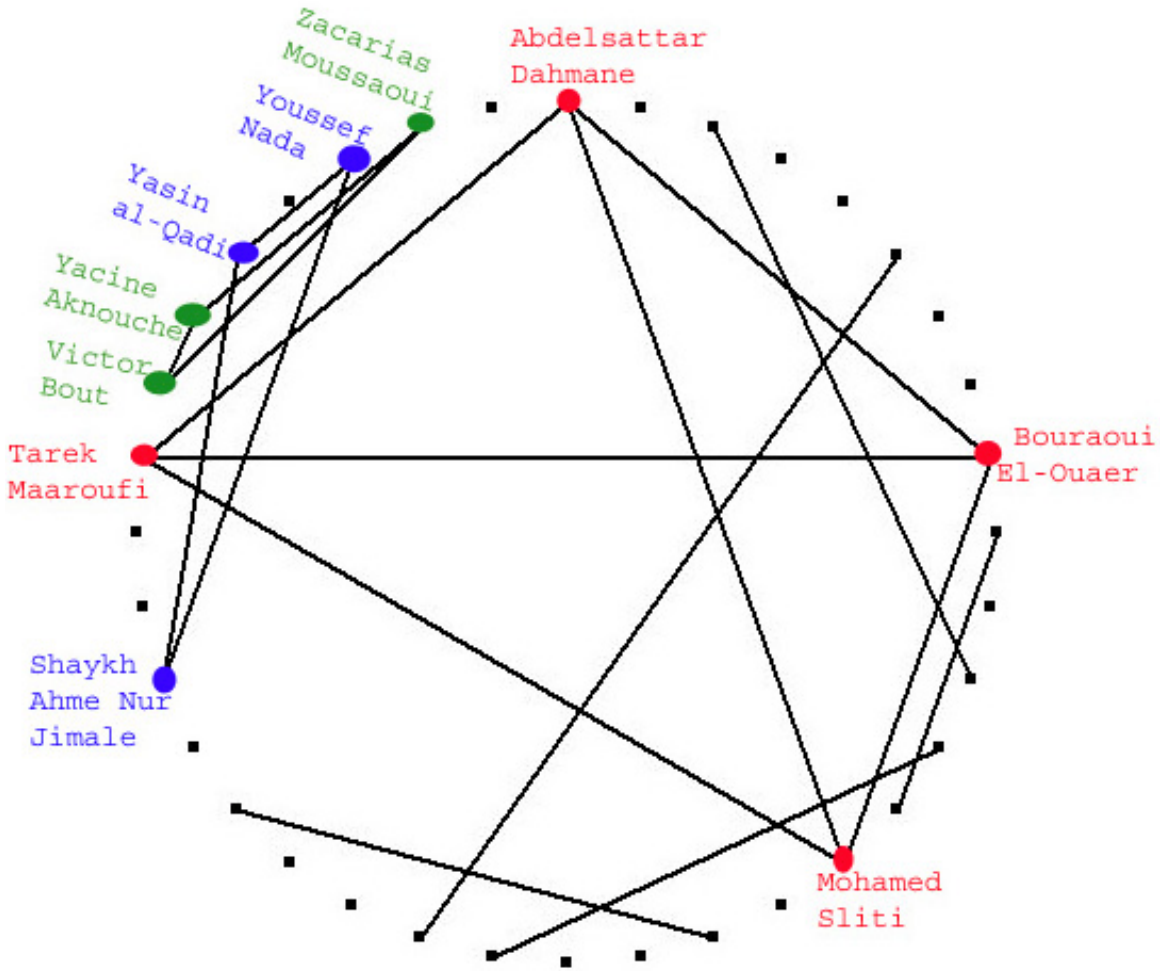
Highly Associated People in Documents

Luis Rocha
2002

rocha@lanl.gov

People Network

Detail



Luis Rocha
2002

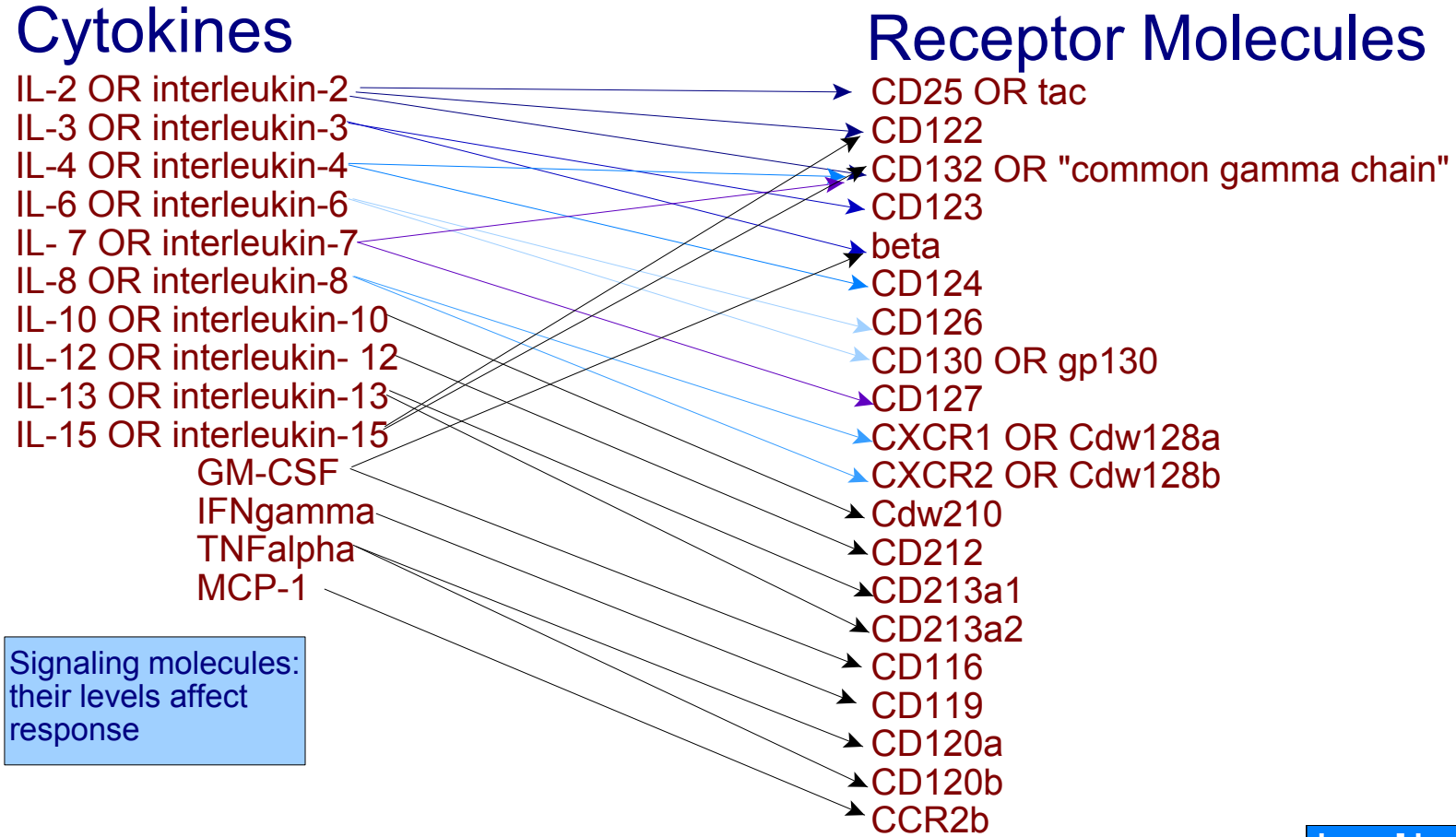
http://www.c3.lanl.gov/~rocha

Los Alamos
National Laboratory

rocha@lanl.gov

Using Information Retrieval

Immunology Testcase



Signaling molecules: their levels affect response

Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Pointwise Mutual Information

Information Retrieval

Turney, P.D. (2001). "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL." *Proceedings of the Twelfth European Conference on Machine Learning* (ECML-2001), Freiburg, Germany, pp. 491-502. http://extractor.iit.nrc.ca/reports/ECML2001.html

- PMI-IR developed to deal with Synonyms
  - Problem word and set of choice words for synonyms (on TOEFL and ESL)
    - Problem: ***levied***
    - Choices: *imposed*, *believed*, *requested*,*correlated*
  - Assigns a *co-occurrence* score to each choice, and selects the choice that maximizes the score.
    - $\text{score}(\text{choice}_i) = p(\text{problem} \mid \text{choice}_i) = p(\text{problem AND choice}_i) / p(\text{choice}_i)$
    - In PMI-IR, the probabilities are calculated using IR from such sources as Altavista.

rocha@lanl.gov

PMI-IR

4 Scores using Altavista

1. Conditional Probability with Altavista AND (when both terms appear in the same document)

$$score_1(choice_i) = \frac{hits(problem \text{ AND } choice_i)}{hits(choice_i)}$$

2. Conditional probability with Altavista NEAR (when both terms appear within 10 words of each other in the same document)

$$score_2(choice_i) = \frac{hits(problem \text{ NEAR } choice_i)}{hits(choice_i)}$$

3. Tends to reduce equal scores for synonyms and antonyms

$$score_3(choice_i) = \frac{hits(problem \text{ NEAR } choice_i) \text{ AND NOT}((problem \text{ OR } choice_i) \text{ NEAR "not"})}{hits(choice_i \text{ AND NOT}(choice_i \text{ NEAR "not"}))}$$

4. Accounts for context words

LSA Problems: "Every year in the early spring farmers [tap] maple syrup from their trees (drain; boil; knock; rap)." The problem word *tap*, out of context, might seem to best match the choice words *knock* or *rap*, but the context maple syrup makes *drain* a better match for tap.

$$score_4(choice_i) = \frac{hits((problem \text{ NEAR } choice_i) \text{ AND } context \text{ AND NOT}((problem \text{ OR } choice_i) \text{ NEAR "not"}))}{hits(choice_i \text{ AND } context \text{ AND NOT}(choice_i \text{ NEAR "not"}))}$$

$$context = \{context_1, context_2, \ldots, context_m\}$$

→ We use NEAR

Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Proximity Queries

$$prox_1(problem, choice_i) = \frac{hits(problem \text{ AND } choice_i)}{hits(problem \text{ OR } choice_i)}$$

$$prox_2(problem, choice_i) = \frac{hits(problem \text{ NEAR } choice_i)}{hits(problem \text{ OR } choice_i)}$$

NEAR: co-occurrence within 10 words

$$prox_3(problem, choice_i) = \frac{hits((problem \text{ NEAR } choice_i) \text{ NEAR } context)}{hits(problem \text{ OR } choice_i)}$$

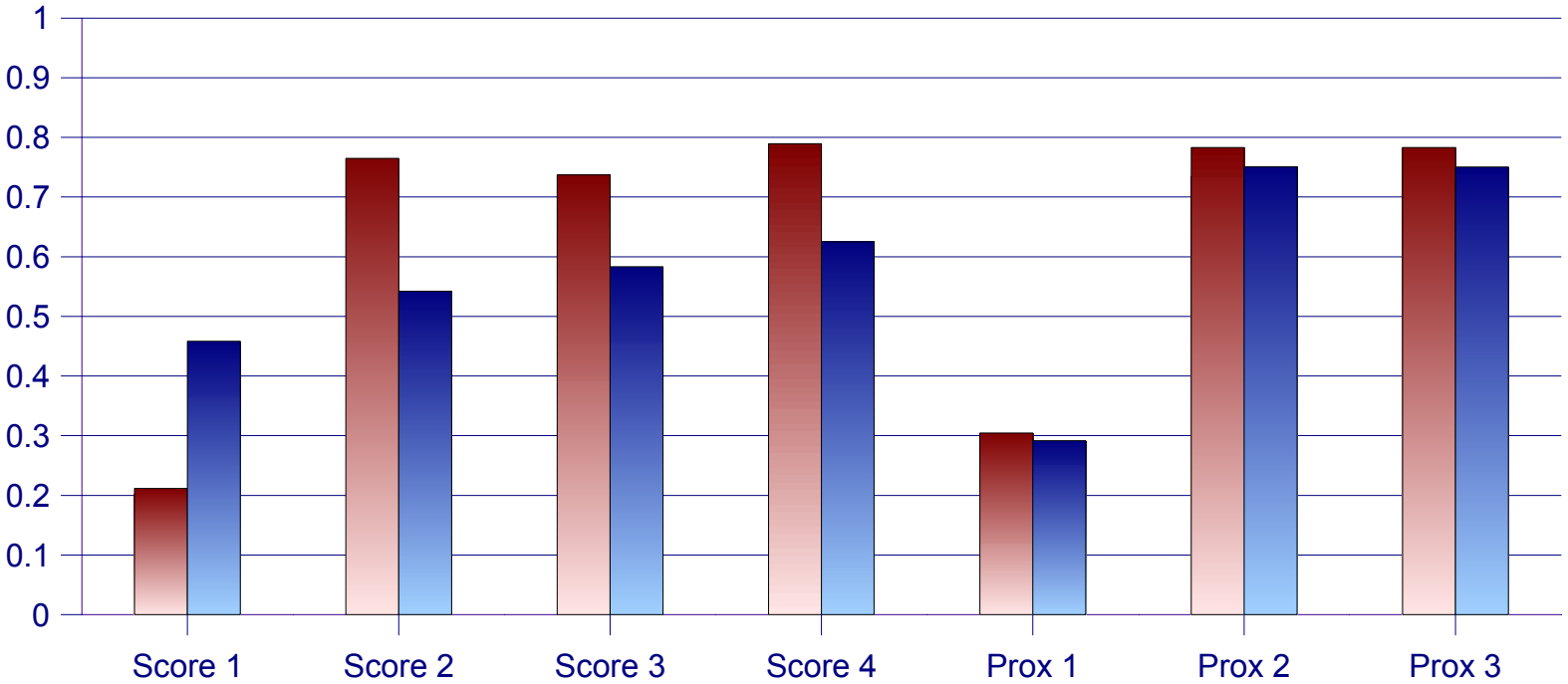
$context = \{\text{receptor}\}$

Luis Rocha
2002

rocha@lanl.gov

Precision and Recall

Values Obtained with FaCSO



Precision: probability that an identified association is relevant

$$precision = \frac{\|\{relevant\} \cap \{retrieved\}\|}{\|\{retrieved\}\|}$$

Recall: probability that an association has been identified given that it is relevant

$$recall = \frac{\|\{retrieved\} \cap \{relevant\}\|}{\|\{relevant\}\|}$$

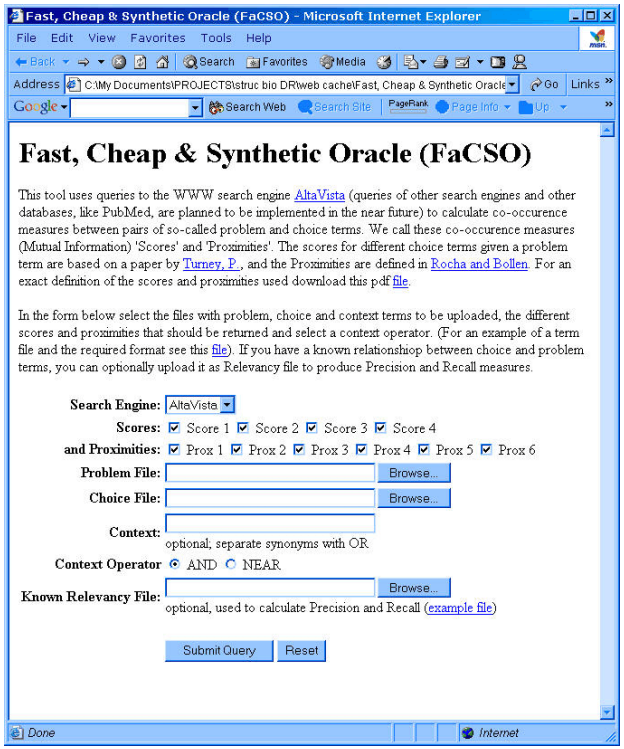
Luis Rocha
2002
http://www.c3.lanl.gov/~rocha

Los Alamos
National Laboratory

rocha@lanl.gov

Measuring Associations

Using Altavista



- Discovers Relevant Associations
- Retrieves Documents substantiating the associations
- Being Developed for PubMed

Fast, Cheap & Synthetic Oracle (FaCSO)

http://bimmer.c3.lanl.gov/~andreas/easyfast8.html

Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

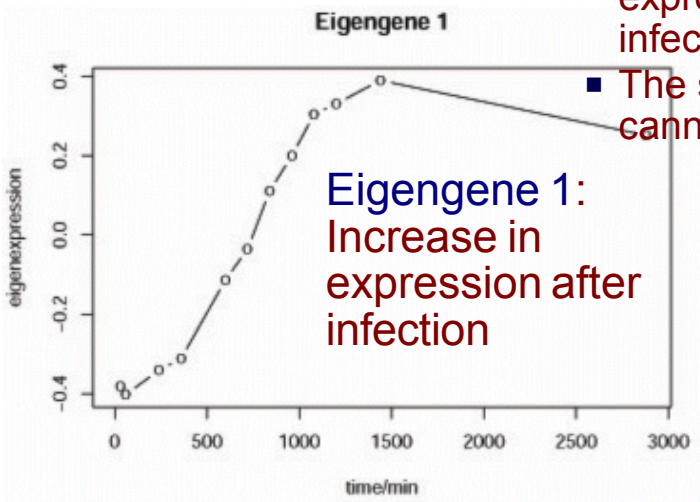
Cytokine Set

Los Alamos National Laboratory

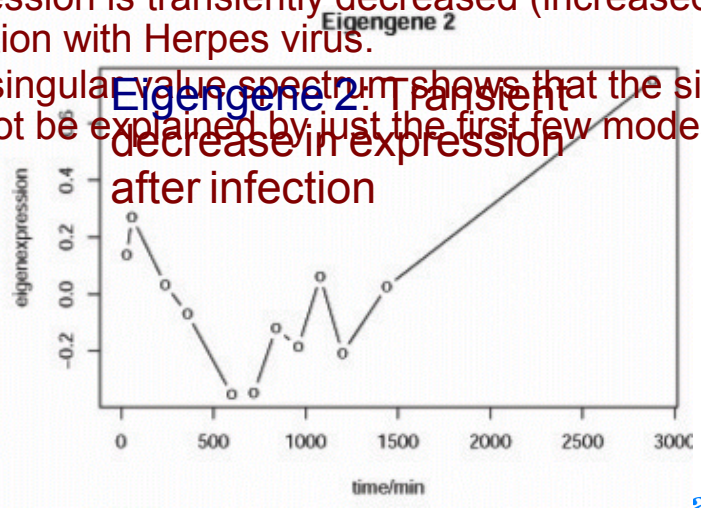
rocha@lanl.gov

SVD of Time-Dependent Expression Data

Gene expression (13000 genes) after infection with Herpes virus

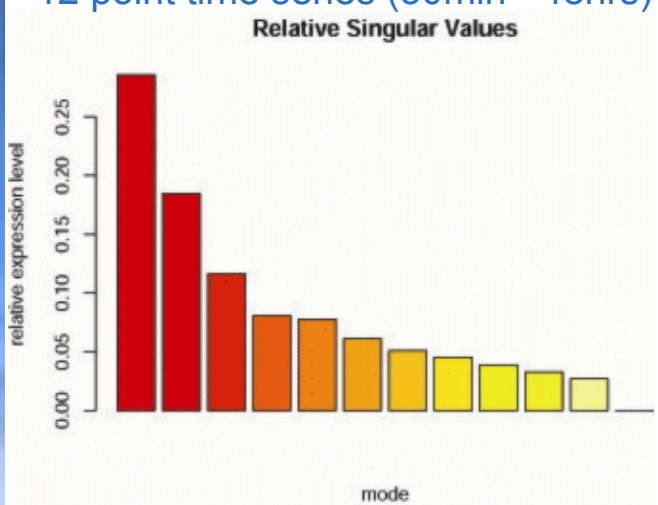


Eigengene 1: Increase in expression after infection

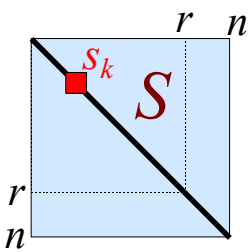


Eigengene 2: Transient decrease in expression after infection

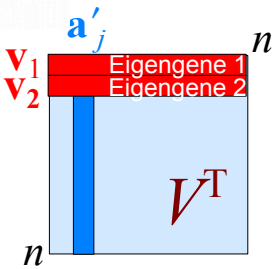
12 point time series (30min - 48hrs)



- Genes whose expression is positively (negatively) correlated with Eigengene 1 are genes whose expression is increased (decreased) after infection with Herpes virus.
- Genes whose expression is positively (negatively) correlated with Eigengene 2 are genes whose expression is transiently decreased (increased) after infection with Herpes virus.
- The singular value spectrum shows that the signal cannot be explained by just the first few modes



$$X = USV^{\mathrm{T}}$$



With Tom Brettin, Michael Wall, Jean Challacombe at LANL and Princeton group (Shenk's lab)

Los Alamos National Laboratory

Luis Rocha
2002

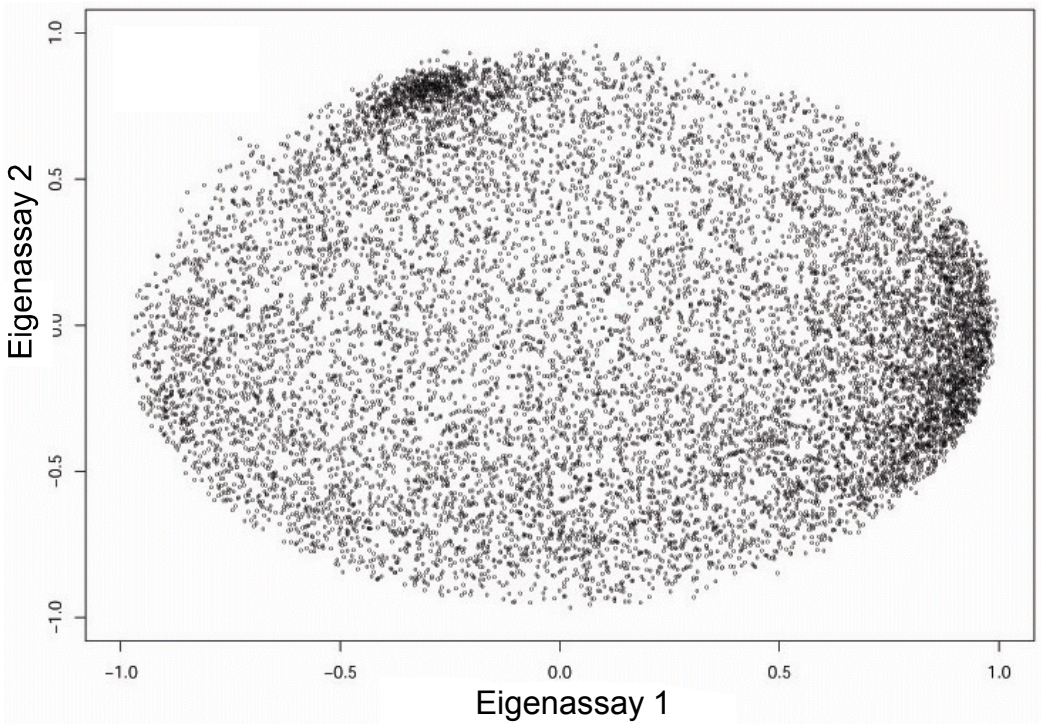
http://www.c3.lanl.gov/~rocha

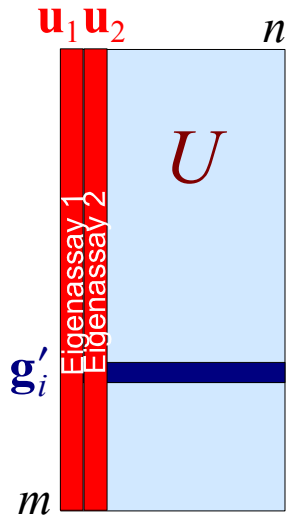
rocha@lanl.gov

Biological Discovery via SVD

Eigenassay Coefficient Plot

LANL group found a second expression mode with interesting biological associations





Princeton group (Shenk's lab) found ~1200 genes that showed significant changes in expression

at least 3 fold change in expression at at least 2 consecutive time points

How to help biologists discover function?

Luis Rocha
2002

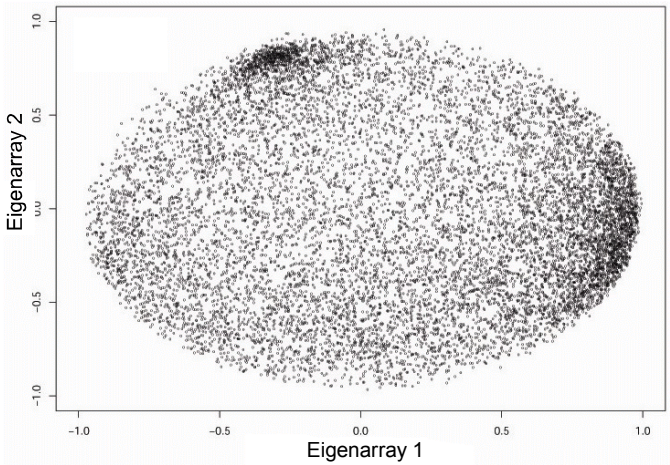
http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Automatic Functional Annotation

Of Gene Clusters



Extract

Sets of genes most correlated and anti-correlated with each eigenarray (or cluster)

Checked likelihood that they are gene symbols (whether or not they co-occur with words like DNA, gene, etc.)

Genbank IDs
Gene symbols

Medline Docs

Extract

MeSH Headings

Count number of times each gene *in a target group* is associated with a heading

Count number of times each gene *in HUGO* is associated with a heading

Statistical measure to see if a heading is mentioned by an unusually large number of genes in target group (given how many HUGO genes mention them)

Collaboration with **Lada Adamic**, **Eytan Adair**, and **Bernardo Huberman** at *HP Labs*, Palo Alto.

Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

MeSH Heading	Sig	genes	ocurrence in target	Occurrence in HUGO
Aspergillosis, Allergic Bronchopulmonary	6.21	CCR2--CR1--HLAA	3	42
Immune Complex Diseases	4.56	CD48--CCR2--CR1--HLAA	4	118
Leukemia, Lymphocytic, Acute, L2	4.29	CR1--FHIT--HLAA--XRCC4	4	129
Sequence Tagged Sites	4.17	CCR2--DAXX--EVI2A--FHIT--GHRHR--HLAA--K	13	924
Leukemia, Myeloid, Chronic-Phase	4.15	CR1--FHIT--HLAA	3	82
Adenocarcinoma, Bronchiolo-Alveolar	3.9	FHIT--HLAA--MUC5B	3	90
Ethers	3.88	CD48--CR1--HLAA--LY6E	4	149
Antibodies, Blocking	3.66	CD48--CCR2--CR1--GHRHR--HLAA--MCAM--C	7	404
Leishmaniasis	3.58	CR1--HLAA--TNP1	3	102
Fructose-Bisphosphatase	3.56	CCR2--FBP2--GIT2	3	103
Bone Marrow Purging	3.4	CD48--CR1--HLAA	3	110
Rhinitis	3.35	CD48--CR1--MUC5B	3	112
Agglutinins	3.35	CR1--HLAA--MUC5B	3	112
Homosexuality	3.25	CCR2--CR1--HLAA	3	117
Linkage (Genetics)	3.22	AMPD3--CAMK2A--RUNX2--CD48--CHRNA3--(	30	3753
Neural Tube Defects	3.18	CD48--GLI3--HLAA--FBP2	4	194
Carcinoid Tumor	3.16	FHIT--GHRHR--HLAA--MUC5B	4	196
Hispanic Americans	3.11	CCR2--CR1--HLAA	3	124
Polymorphism, Restriction Fragment Lengt	3.1	CAMK2A--CCR2--CR1--EVI2A--FHIT--GHRHR-	21	2355
Communicable Diseases	3.09	CD48--CR1--HLAA	3	125
AIDS Vaccines	3.06	CCR2--CR1--HLAA	3	127
Gene Order	3.06	HLAA--TNP1--SNX3	3	127
Tuberculosis	3.04	CD48--CCR2--CR1--HLAA--TNP1	5	293
Hemoglobinuria, Paroxysmal	3.04	CD48--CR1--HLAA	3	128
Lupus Nephritis	3.02	CD48--CCR2--CR1--HLAA	4	207
Hirschsprung Disease	2.99	EDN3--GLI3--HLAA	3	131
Belgium	2.93	CCR2--CR1--HLAA	3	134
Protein Isoforms	2.9	ABCA2--RUNX2--CD48--CCR2--CR1--GHRHR--	20	2298
Haplotypes	2.88	CD48--CCR2--CR1--EDN3--EVI2A--GHRHR--HI	16	1700

rocha@lanl.gov

MeSH Headings

Correlated with Second Eigenassay

Genes involved in transcription regulation, immune response, oncogenesis as well as growth factors/cytokines and their receptors

MeSH Heading	Sig	genes	ocurrence in target	Occurrence in HUGO
Aspergillosis, Allergic Bronchopulmonary	6.21	CCR2--CR1--HLAA	3	42
Immune Complex Diseases	4.56	CD48--CCR2--CR1--HLAA	4	118
Leukemia, Lymphocytic, Acute, L2	4.29	CR1--FHIT--HLAA--XRCC4	4	129
Sequence Tagged Sites	4.17	CCR2--DAXX--EVI2A--FHIT--GHRHR--HLAA--K	13	924
Leukemia, Myeloid, Chronic-Phase	4.15	CR1--FHIT--HLAA	3	82
Adenocarcinoma, Bronchiolo-Alveolar	3.9	FHIT--HLAA--MUC5B	3	90
Ethers	3.88	CD48--CR1--HLAA--LY6E	4	149
Antibodies, Blocking	3.66	CD48--CCR2--CR1--GHRHR--HLAA--MCAM--C	7	404
Leishmaniasis	3.58	CR1--HLAA--TNP1	3	102
Fructose-Bisphosphatase	3.56	CCR2--FBP2--GIT2	3	103
Bone Marrow Purging	3.4	CD48--CR1--HLAA	3	110
Rhinitis	3.35	CD48--CR1--MUC5B	3	112
Agglutinins	3.35	CR1--HLAA--MUC5B	3	112
Homosexuality	3.25	CCR2--CR1--HLAA	3	117
Linkage (Genetics)	3.22	AMPD3--CAMK2A--RUNX2--CD48--CHRNA3--(	30	3753
Neural Tube Defects	3.18	CD48--GLI3--HLAA--FBP2	4	194
Carcinoid Tumor	3.16	FHIT--GHRHR--HLAA--MUC5B	4	196
Hispanic Americans	3.11	CCR2--CR1--HLAA	3	124
Polymorphism, Restriction Fragment Lengt	3.1	CAMK2A--CCR2--CR1--EVI2A--FHIT--GHRHR-	21	2355
Communicable Diseases	3.09	CD48--CR1--HLAA	3	125
AIDS Vaccines	3.06	CCR2--CR1--HLAA	3	127
Gene Order	3.06	HLAA--TNP1--SNX3	3	127
Tuberculosis	3.04	CD48--CCR2--CR1--HLAA--TNP1	5	293
Hemoglobinuria, Paroxysmal	3.04	CD48--CR1--HLAA	3	128
Lupus Nephritis	3.02	CD48--CCR2--CR1--HLAA	4	207
Hirschsprung Disease	2.99	EDN3--GLI3--HLAA	3	131
Belgium	2.93	CCR2--CR1--HLAA	3	134
Protein Isoforms	2.9	ABCA2--RUNX2--CD48--CCR2--CR1--GHRHR-	20	2298
Haplotypes	2.88	CD48--CCR2--CR1--EDN3--EVI2A--GHRHR--H	16	1700

Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Distance from Proximity

Semi-metric Behavior

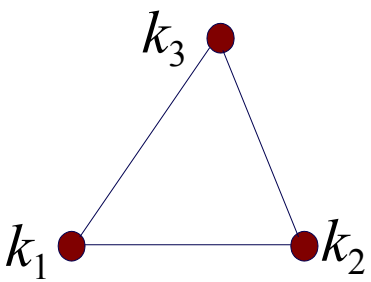
$$kdp(k_i,k_j)=\frac{\sum_{k=1}^{m}(a_{i,k}\wedge a_{j,k})}{\sum_{k=1}^{m}(a_{i,k}\vee a_{j,k})}=\frac{N_{\cap}(k_i,k_j)}{N_{\cup}(k_i,k_j)}$$

(Keyword Document Proximity)

$$d_{kdp}(k_i,k_j)=\frac{1}{kdp(k_i,k_j)}-1$$

(Keyword Document Distance)

d is a distance function because it is a nonnegative, symmetric, real-valued function such that $d(k, k) = 0$

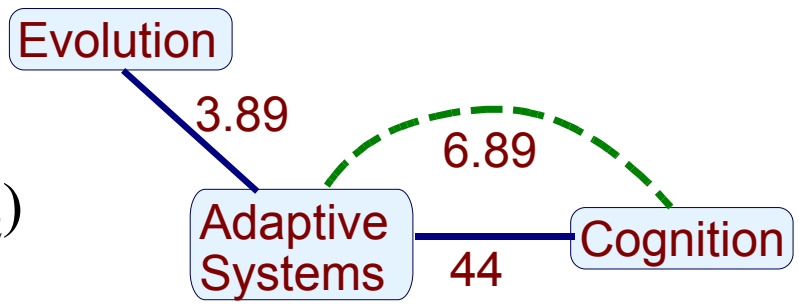


$$d(k_1, k_2) \le d(k_1, k_3) + d(k_3, k_2)$$

Metric

$$d(k_1, k_2) > d(k_1, k_3) + d(k_3, k_2)$$

Semi-metric



Semi-metric ratio: 6.3861

Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Distance from Proximity

Generic Case

$$XYP\left(x_i, x_j\right) = \frac{\sum_{k=1}^{m}\left(r_{i,k} \wedge r_{j,k}\right)}{\sum_{k=1}^{m}\left(r_{i,k} \vee r_{j,k}\right)}; \quad YXP\left(y_i, y_j\right) = \frac{\sum_{k=1}^{n}\left(r_{k,i} \wedge r_{k,j}\right)}{\sum_{k=1}^{n}\left(r_{k,i} \vee r_{k,j}\right)}$$

$$d_X\left(x_i, x_j\right) = \frac{1}{XYP\left(x_i, x_j\right)} - 1; \quad d_Y\left(y_i, y_j\right) = \frac{1}{YXP\left(y_i, y_j\right)} - 1$$

Distance from a Proximity Graph is semi-Metric
Distance from a Similarity Graph is Metric

Luis Rocha
2002

Los Alamos
National Laboratory

rocha@lanl.gov

Measuring Semi-Metric Behavior

Semi-metric Measures

- **Semi-metric ratio**
  - Absolute measure of indirect distance reduction

$$s(x_i, x_j) = \frac{d_{direct}(x_i, x_j)}{d_{shortest}(x_i, x_j)}$$

- **Relative Semi-metric ratio**
  - Distance reduction against maximum contraction

$$rs(x_i, x_j) = \frac{d_{direct}(x_i, x_j) - d_{shortest}(x_i, x_j)}{d_{\max} - d_{\min}}$$

- **Below Average Ratio**
  - Captures semi-metric distance reductions which contract to below the average distance for a given node. Captures some of the cases of initial ∞ distance

$$b(x_i, x_j) = \frac{\overline{d_{x_i}}}{d_{shortest}(x_i, x_j)}$$

Luis Rocha
2002

rocha@lanl.gov

User Interests

86 Keywords

Dissertation Database

ADAPTIVE SYSTEMS	COGNITION	Semi-metric Ratio	6.3861	Relative Semi-metric	0.8434	1
EVOLUTION	CONSTRUCTIVISM	Semi-metric Ratio	5.0000	Relative Semi-metric	0.7636	2
EVOLUTION	PSYCHOLOGY	Semi-metric Ratio	5.0000	Relative Semi-metric	0.7273	3
EVOLUTION	DNA	Semi-metric Ratio	4.6936	Relative Semi-metric	0.6439	4
LIFE	COGNITION	Semi-metric Ratio	4.5455	Relative Semi-metric	0.6559	5
EVOLUTION	CONTROL	Semi-metric Ratio	4.5407	Relative Semi-metric	0.7620	6
MATHEMATICS	SYSTEMS THEORY	Semi-metric Ratio	4.4205	Relative Semi-metric	0.6507	7
CYBERNETICS	THEORETICAL BIOLOGY	Semi-metric Ratio	4.3593	Relative Semi-metric	0.5780	8
ARTIFICIAL LIFE	NEURAL NETWORKS	Semi-metric Ratio	4.2918	Relative Semi-metric	0.3486	9
ARTIFICIAL INTELLIGENCE	SELF-ORGANIZING SYSTEMS	Semi-metric Ratio	4.0897	Relative Semi-metric	0.5323	10
ROBOTICS	COGNITION	Semi-metric Ratio	3.9759	Relative Semi-metric	0.5614	11
MATHEMATICS	MEMORY	Semi-metric Ratio	3.8509	Relative Semi-metric	0.5216	12
ROBOTICS	MATHEMATICS	Semi-metric Ratio	3.8028	Relative Semi-metric	0.4523	13
LIFE	PHILOSOPHY	Semi-metric Ratio	3.7928	Relative Semi-metric	0.6861	14
ROBOTICS	PHILOSOPHY	Semi-metric Ratio	3.7525	Relative Semi-metric	0.6168	15
PHILOSOPHY	EMERGENCE	Semi-metric Ratio	3.7365	Relative Semi-metric	0.6325	16
COGNITION	SEMIOTICS	Semi-metric Ratio	3.7353	Relative Semi-metric	0.5825	17
EVOLUTION	MOLECULAR BIOLOGY	Semi-metric Ratio	3.7229	Relative Semi-metric	0.5984	18
EVOLUTION	CELLULAR AUTOMATA	Semi-metric Ratio	3.7218	Relative Semi-metric	0.6316	19
EVOLUTION	INFORMATION THEORY	Semi-metric Ratio	3.7037	Relative Semi-metric	0.7134	20

9% Semi-metric
25% below Average

COMPUTER NETWORKS	STATE-SPACE METHODS	Under Average	7.5000
MORPHOLOGY	GROWTH	Under Average	5.3000
COMPUTER GRAPHICS	FORM	Under Average	4.7037
MORPHOLOGY	FORM	Under Average	4.2400
MORPHOLOGY	PALEONTOLOGY	Under Average	3.5333
MORPHOLOGY	CONTINGENCY	Under Average	3.5333
ROBOTICS	NEURAL NETWORKS	Under Average	3.3420
EVOLUTION	NEURAL NETWORKS	Under Average	3.1134
EVOLUTION	DNA	Under Average	3.0404
COGNITION	THEORETICAL BIOLOGY	Under Average	3.0377

Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Trends in Collections

500 Keywords

ARP Database

leukemia	myocardi	Semi-metric Ratio	272.1996	Relative Semi-metric	0.4981	1
hormon	thin	Semi-metric Ratio	214.0797	Relative Semi-metric	0.9953	2
care	excit	Semi-metric Ratio	213.5900	Relative Semi-metric	0.9953	3
gene	equat	Semi-metric Ratio	205.7649	Relative Semi-metric	0.9951	4
film	transcript	Semi-metric Ratio	204.5103	Relative Semi-metric	0.9951	5
spectroscopi	care	Semi-metric Ratio	194.3478	Relative Semi-metric	0.9949	6
transcript	thin	Semi-metric Ratio	193.0644	Relative Semi-metric	0.9948	7
pressur	t-cell	Semi-metric Ratio	190.8173	Relative Semi-metric	0.9948	8
film	mutat	Semi-metric Ratio	186.8350	Relative Semi-metric	0.9946	9
vascular	catalyst	Semi-metric Ratio	185.3671	Relative Semi-metric	0.9946	10
film	endoth	Semi-metric Ratio	183.0219	Relative Semi-metric	0.9945	11
film	macrophag	Semi-metric Ratio	180.4128	Relative Semi-metric	0.9945	12
nonlinear	nerv	Semi-metric Ratio	177.6419	Relative Semi-metric	0.9944	13
film	clone	Semi-metric Ratio	175.6775	Relative Semi-metric	0.9943	14
mutat	equat	Semi-metric Ratio	175.1138	Relative Semi-metric	0.9943	15
film	secretion	Semi-metric Ratio	174.9438	Relative Semi-metric	0.9943	16
thin	endoth	Semi-metric Ratio	173.8007	Relative Semi-metric	0.9942	17
pressur	leukemia	Semi-metric Ratio	172.2365	Relative Semi-metric	0.9942	18
thin	macrophag	Semi-metric Ratio	171.4462	Relative Semi-metric	0.9942	19
film	mortal	Semi-metric Ratio	169.9975	Relative Semi-metric	0.9941	20
clone	thin	Semi-metric Ratio	167.1643	Relative Semi-metric	0.9940	21

95% Semi-metric
35% Below Average

leukemia	myocardi	Under Average	77.7665
thin	nitric	Under Average	50.5213
equat	messenger-rna	Under Average	42.3600
chemotherapi	myocardi	Under Average	41.6956
nonlinear	nerv	Under Average	40.1634
film	risk	Under Average	40.0989
equat	transcript	Under Average	39.9494
equat	clone	Under Average	39.9156
film	hormon	Under Average	39.6987
equat	gene-express	Under Average	37.0435

Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

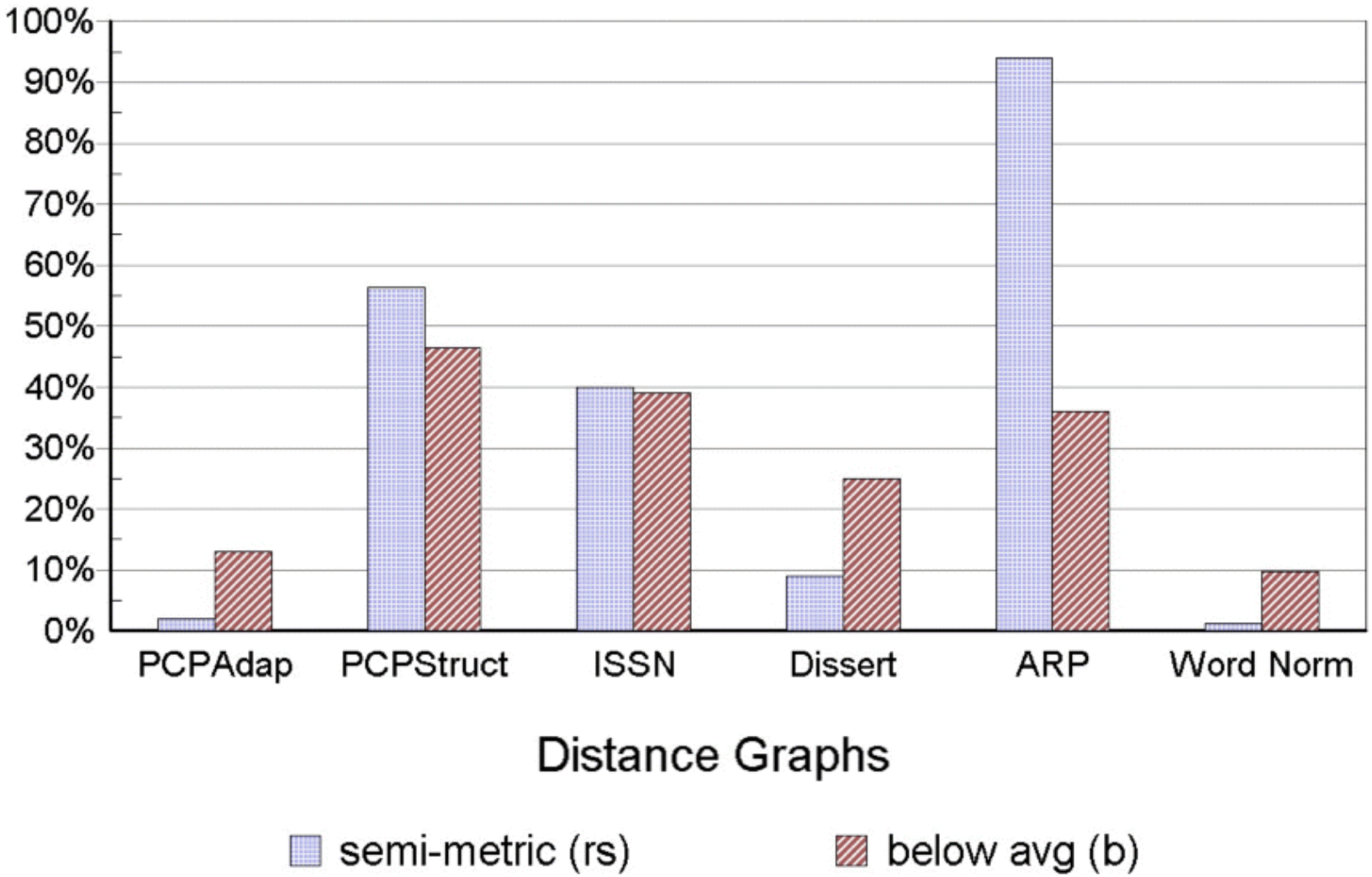
Analysing Different Document Networks

For Semi-Metric Behavior

- PCP Web
  - Collection of dictionary-like definitions about Systems Research topics; each of these 423 web pages is associated with a specific concept (e.g. "Adaptive Systems").
    - Proximity Data extracted from hyperlink structure (Symmetric)
    - Adapted Hyperlink Structure from user paths extracted from web logs using Bollen's Algorithms (Symmetric)
- ISSN
  - Network of 472 Research Journal Titles (e.g. "Communications of the ACM" and "BioSystems"), identified by their ISSN. Adapted using the same methodology used for the adapted PCP web site data (symmetric).
- Word Norm
  - Nelson et al' s associative graphs between pairs of words from free association experiments with more than 6000 subjects; weights characterize the semantic proximity between words as understood by the population of subjects.We used a subset of 150 words from this dataset (of about 5000 words): the 150 most common English nouns, (words such as "art", "car", "face").
- Random Distance Graphs
  - Uniform, Exponential, Hyperexponential Proximity Distributions

rocha@lanl.gov

Percentage of Semi-metric Pairs

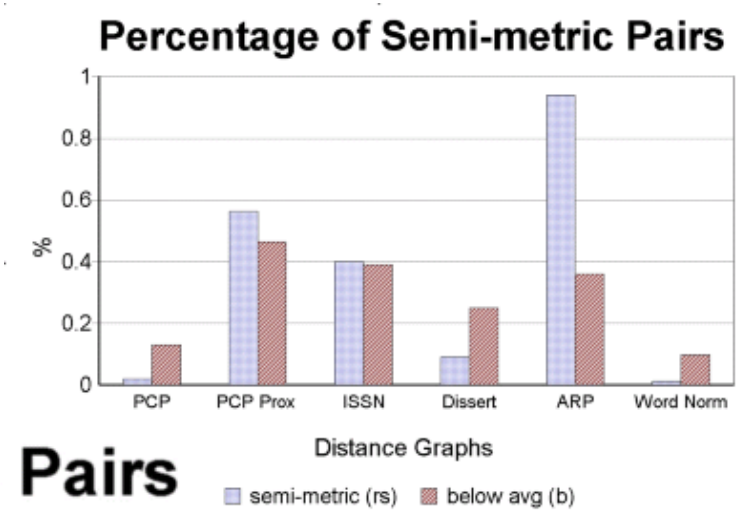


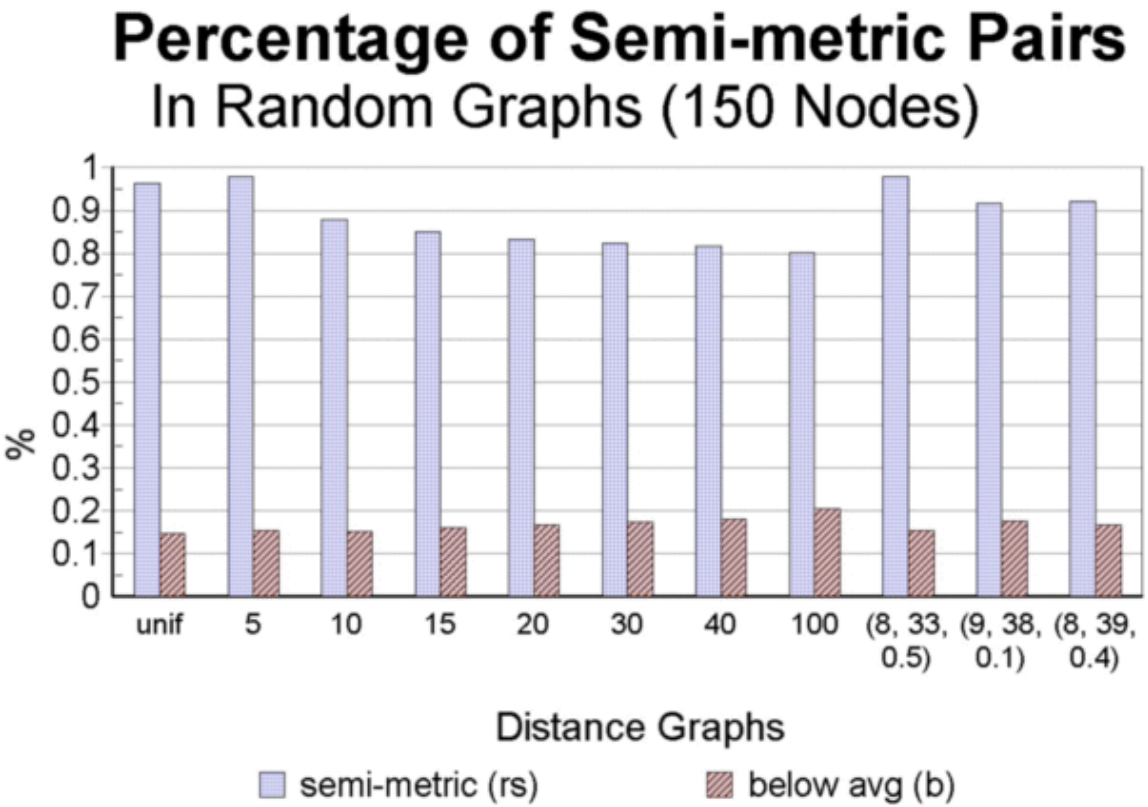
Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos
National Laboratory

rocha@lanl.gov





Luis Rocha
2002

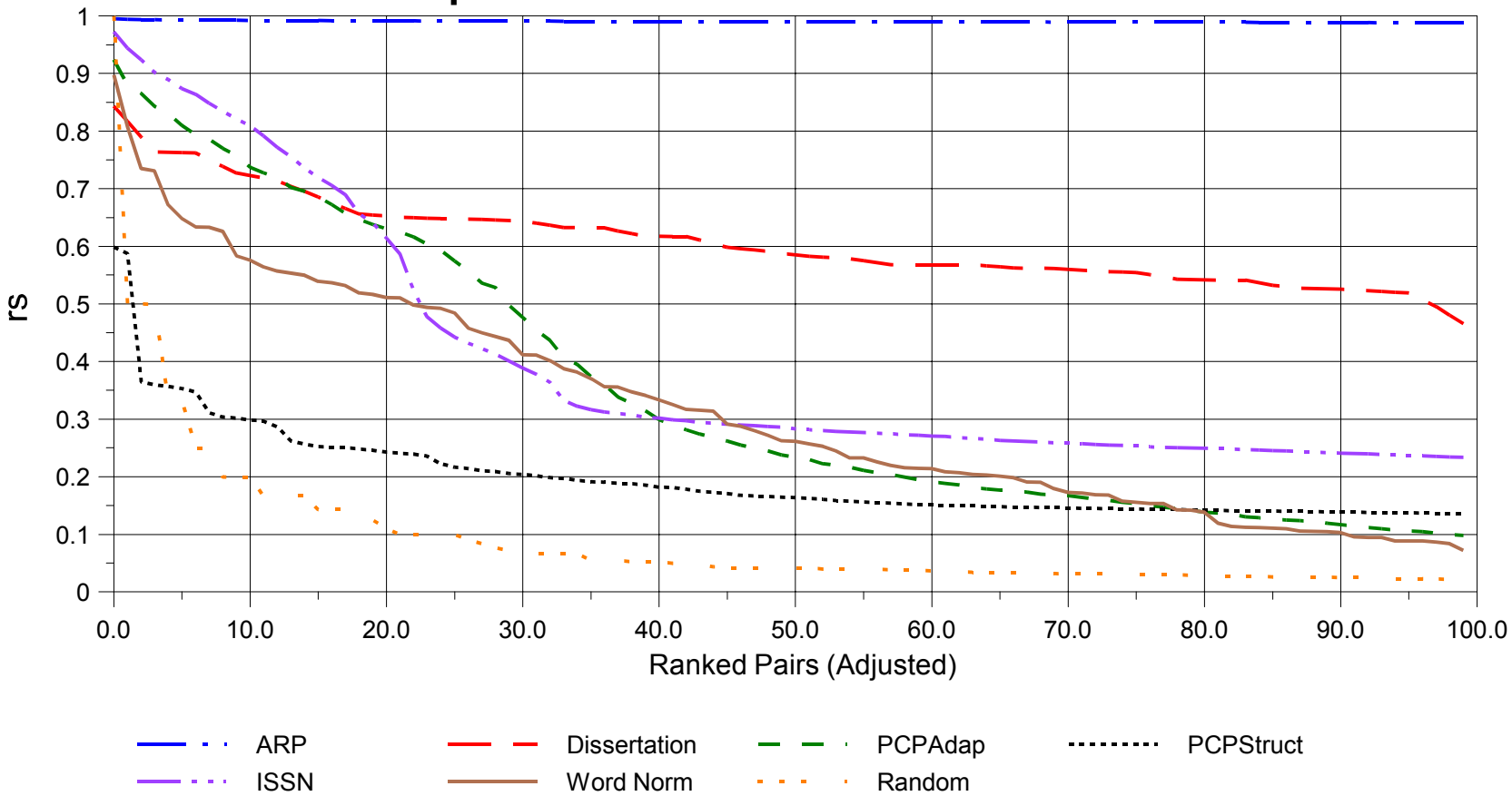
http://www.c3.lanl.gov/~rocha

Los Alamos
National Laboratory

rocha@lanl.gov

TRS$^{1\%}$

Top 1% Semimetric Pairs

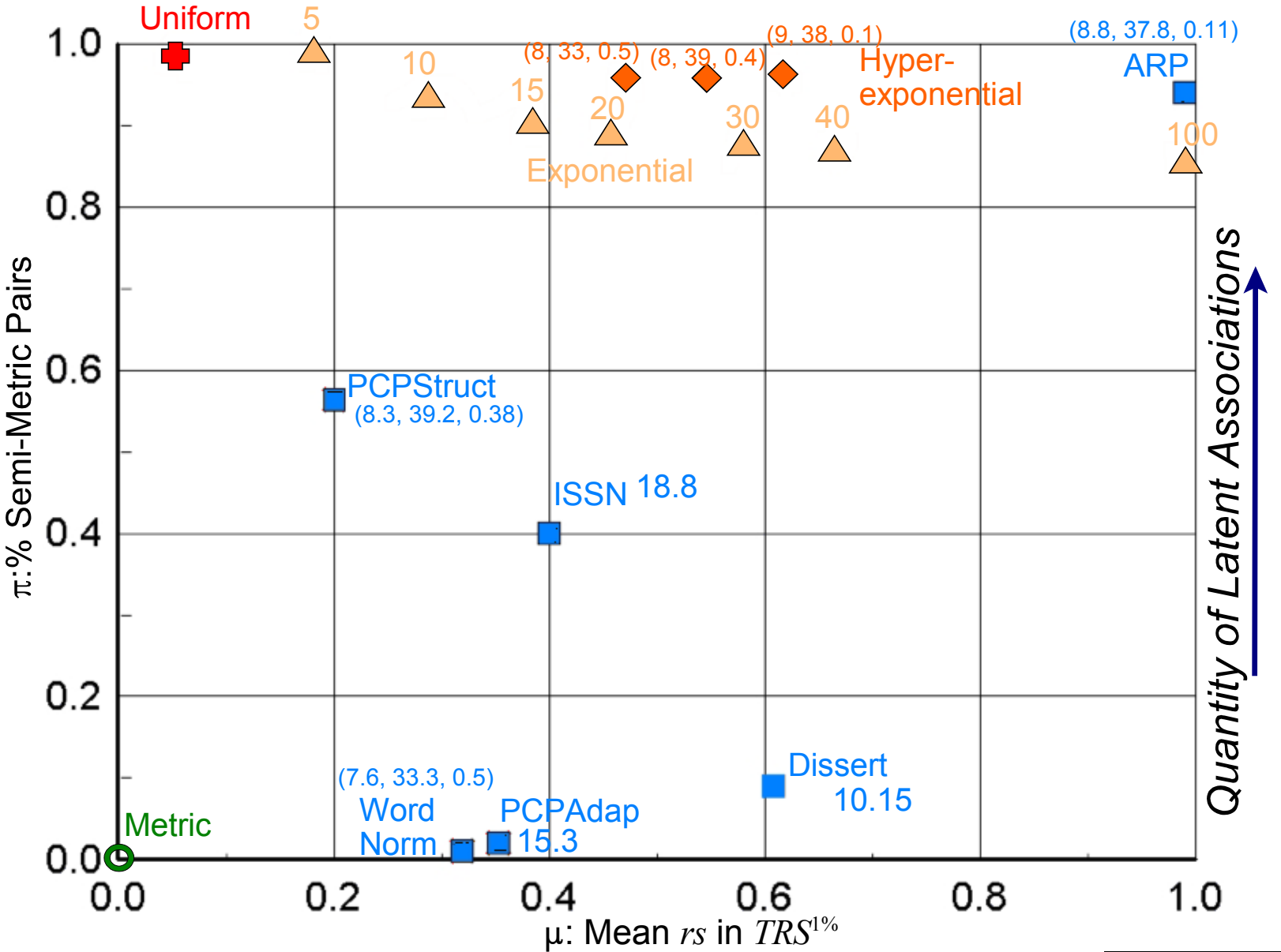


Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov



Luis Rocha
2002

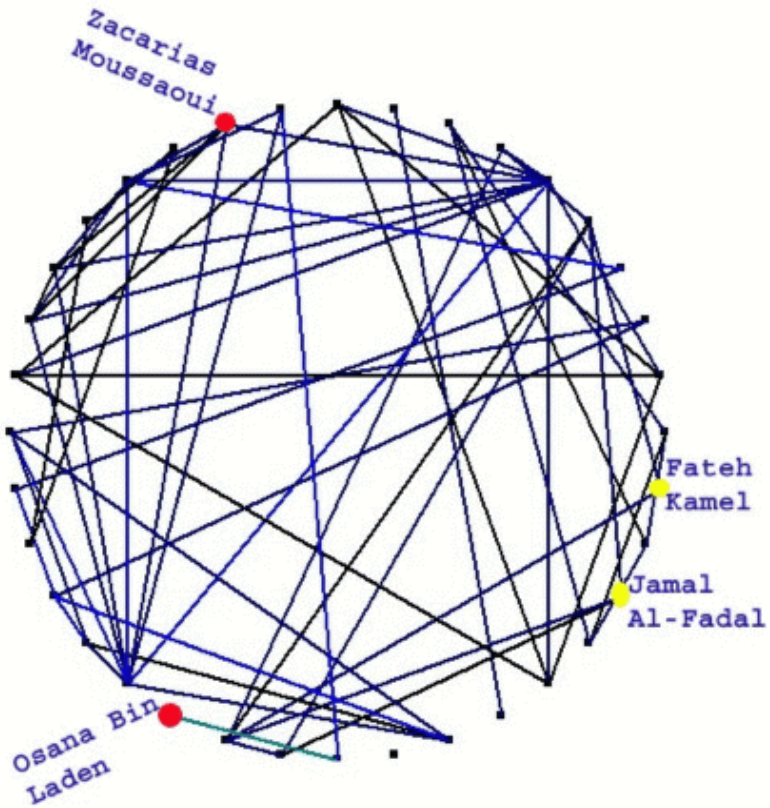
http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Identification of Latent Associations

Terrorist Network Example



- Pairs with larger semi-metric behavior denote a *latent association*
  - Not grounded on direct evidence provided by the relation R, but rather implied by the overall network of associations in this relation.
  - Meaning depends on the semantics of the application
    - In graphs of keyword co-occurrence in documents: associated with novelty and can be used to identify trends.
    - In terrorist networks it may identify pairs of people, groups, etc. for which we do not have direct evidence, in the available documents, that a real association exists, but who could easily be indirectly associated.
  - In recommendation system for journals now at LANL

Luis Rocha
2002

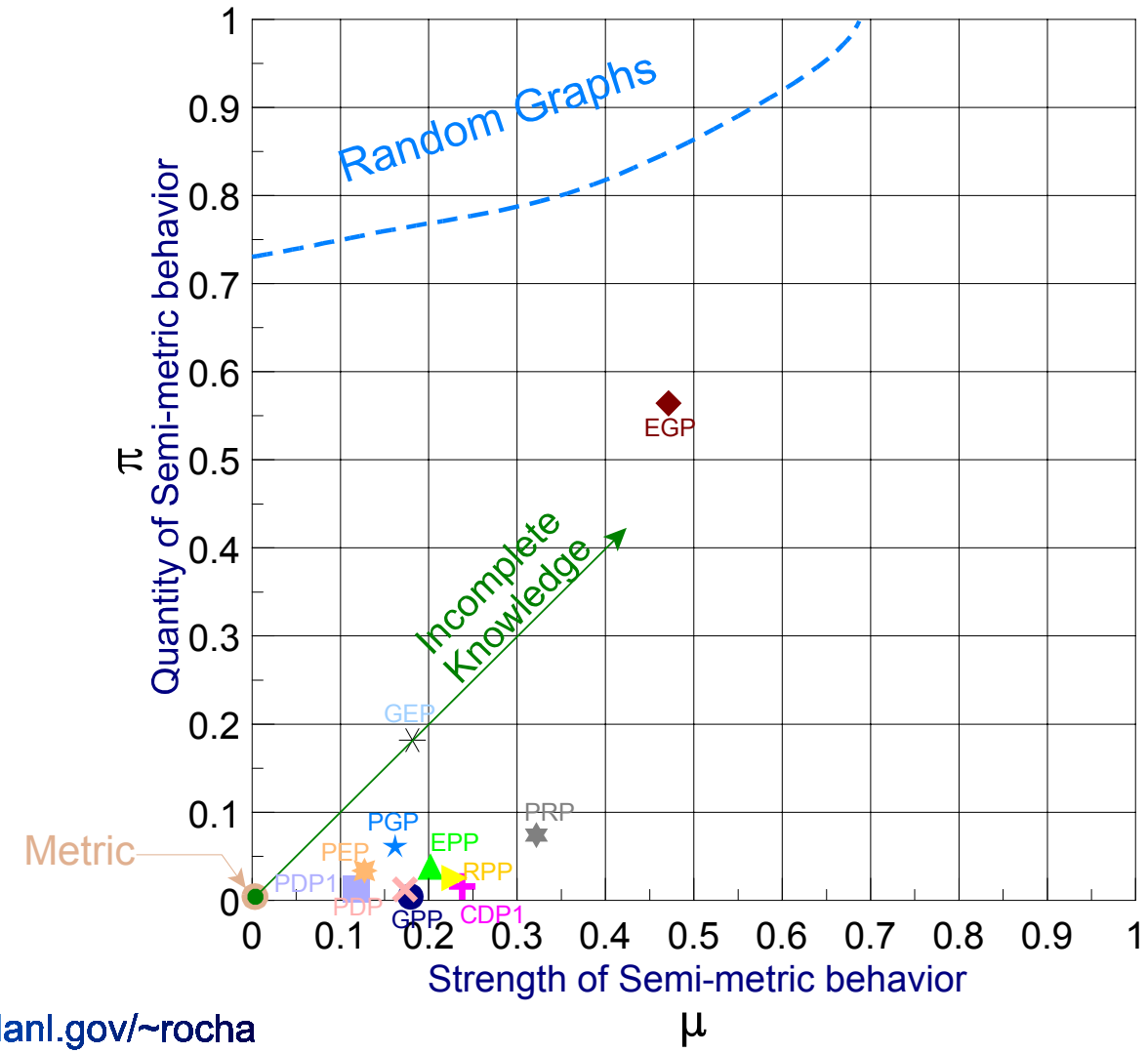
http://www.c3.lanl.gov/~rocha

Los Alamos
National Laboratory

rocha@lanl.gov

Does a latent association imply missing evidence?

Incomplete Knowledge



Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

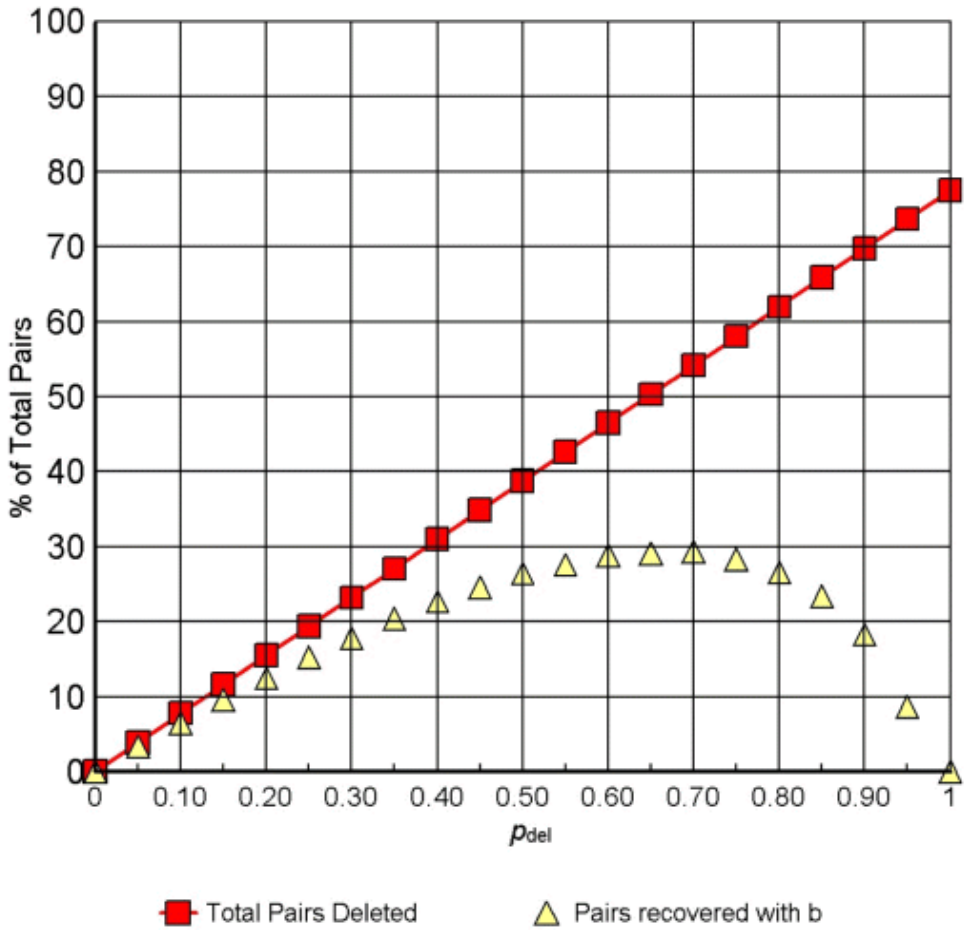
Los Alamos National Laboratory

rocha@lanl.gov

Detecting Incomplete Knowledge

Random Deletion Experiments

- Perfect Knowledge
  - Transitive Closure of real graph
  - Metric Distance Graph
- Incomplete Knowledge
  - Each positive association is deleted with probability p_{del}
  - 100 graphs for each value of p_{del}



Full Deletion

Luis Rocha
2002

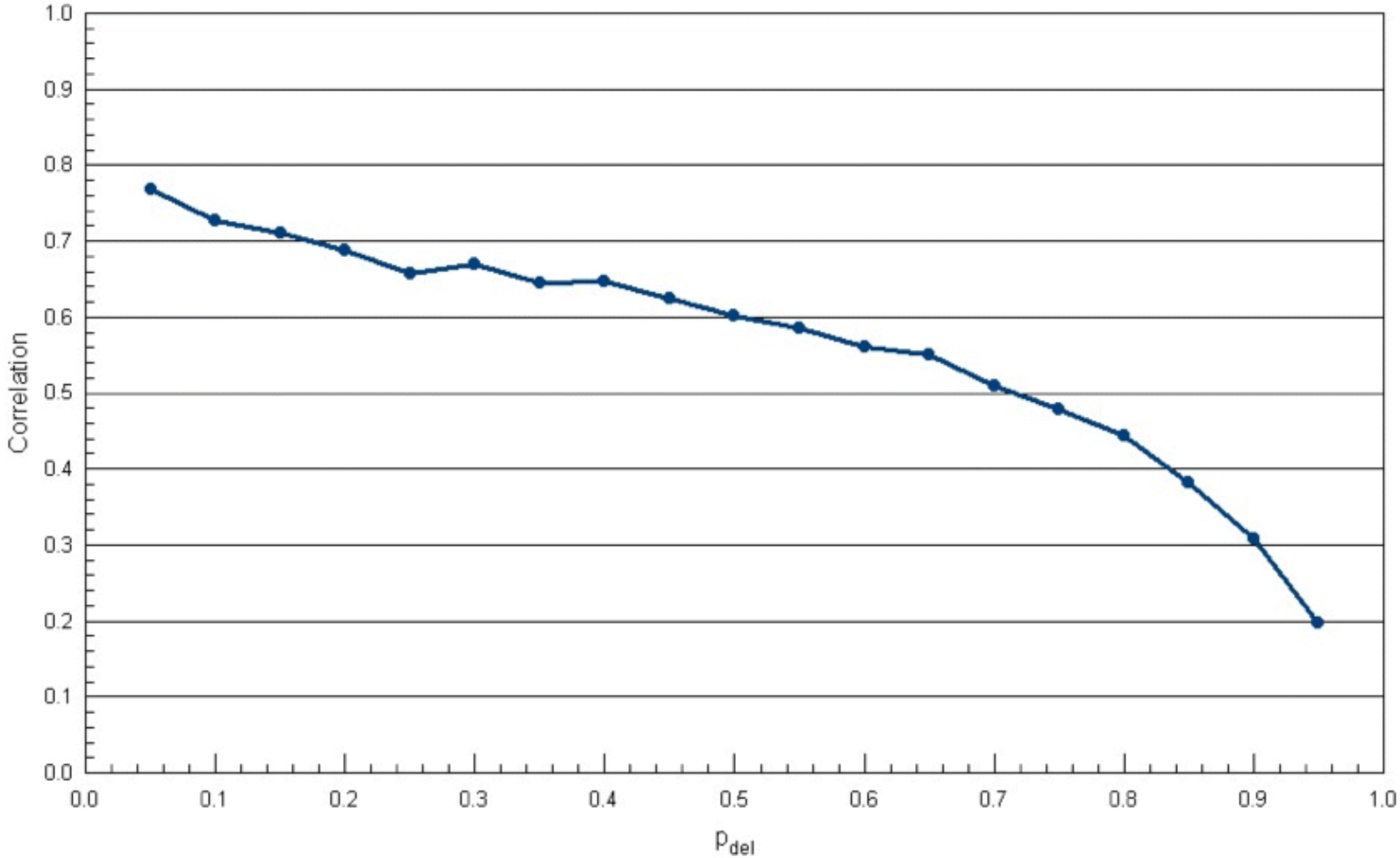
http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Full Deletion

Correlation of value of b with deleted value



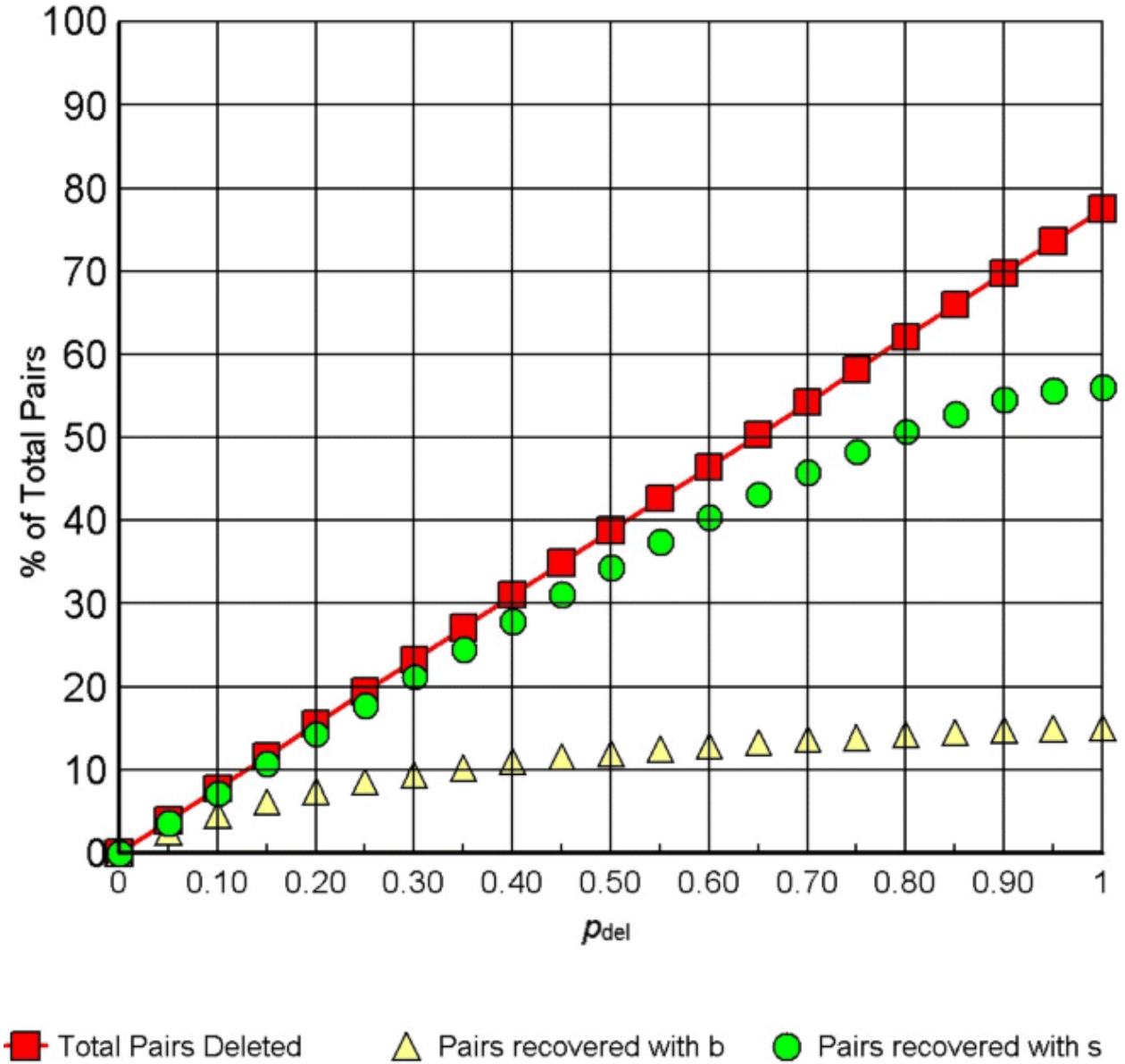
Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Partial Deletion



Luis Rocha
2002

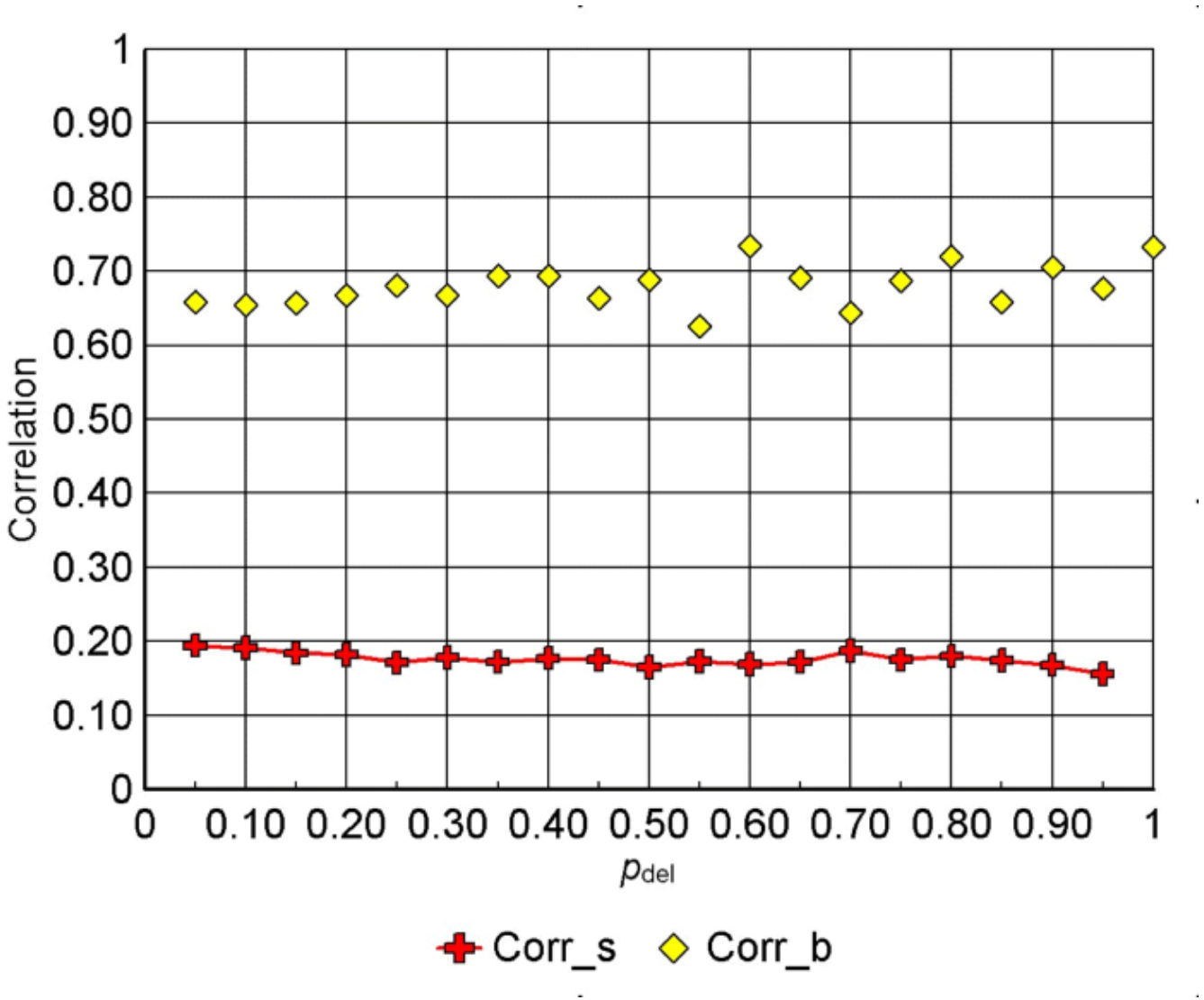
http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Partial Deletion

Parameter Correlation



Luis Rocha
2002

http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

Proximity 6

$$prox_6(problem, choice_i) = \frac{hits((problem\ \text{NEAR}\ choice_i)\ \text{NEAR}\ context)}{hits(problem\ \text{OR}\ choice_i)}$$

Semi-metric Associatons

Cytokines: Problem Terms

	IL-2	IL-3	IL-4	IL-6	IL-7	IL-8	IL-10	IL-12	IL-13	IL-15	GM-CSF	IFNγ	TNFα
CD25	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CD122	0.0008	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000
CD132	0.0005	0.0001	0.0011	0.0000	0.0027	0.0000	0.0000	0.0000	0.0001	0.0025	0.0000	0.0000	0.0000
CD123	0.0001	0.0011	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
beta	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CD124	0.0000	0.0004	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000
CD126	0.0000	0.0000	0.0001	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CD130	0.0001	0.0002	0.0000	0.0067	0.0006	0.0000	0.0001	0.0001	0.0000	0.0001	0.0003	0.0000	0.0000
CD127	0.0000	0.0000	0.0000	0.0002	0.0009	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CXCR1	0.0000	0.0000	0.0000	0.0000	0.0002	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CXCR2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000
Cdw210	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CD212	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
CD213a1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0007	0.0000	0.0000	0.0000	0.0000
CD213a2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0001	0.0000	0.0000	0.0000
CD116	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0012	0.0000	0.0000
CD119	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000
CD120a	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CD120b	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0012
CCR2b	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Receptor Molecules: choice terms

Existing co-occurrence

IL-15, CD122
IL-2, CD130
IL-7, CD126
IL-4, CD122
IL-13, CD132
IL-10, CD130

Null co-occurrence

IL-7, CD122
IL-15, CD123
IL-15, CD127
TNFalpha, CD120a
IL-15, CD124
IL-7, CD123
IL-15, CD120a

http://www.c3.lanl.gov/~rocha

Luis Rocha
2002

Los Alamos National Laboratory